Meta-Album: Datasheets for Datasets

Ihsan Ullah[*], **Dustin Carrión-Ojeda**[*¶‖], **Sergio Escalera**[#%], **Isabelle Guyon**[*%], **Mike Huisman**[+], **Felix Mohr**[‡], **Jan N. van Rijn**[+], **Haozhe Sun**[*], **Joaquin Vanschoren**[§], **Phan Anh Vu**[*]
% ChaLearn, USA
‖ hessian.AI, Germany
Universitat de Barcelona, Spain
‡ Universidad de La Sabana, Colombia
¶ Technische Universität Darmstadt, Germany
∗ LISN/CNRS/INRIA, Université Paris-Saclay, France
§ TU/e Eindhoven University of Technology, The Netherlands
+ Leiden Institute of Advanced Computer Science (LIACS), Leiden University, the Netherlands
https://meta-album.github.io/

Datasheet for LR_AM.BRD Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Dustin Carrión created this dataset under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of birds on natural backgrounds. In the original dataset, the resolution of the instances is 224x224 pixels.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 12600 instances from 315 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images per class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance has a category/label, which is provided with the meta-data. A category corresponds to a bird species.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

In the original dataset, the data (images of birds) were collected from internet searches by species name. For the formatted dataset, the collecting mechanism was sampling from the original dataset.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original author of the Birds dataset: Gerald Piosenka.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

N/A

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The images are resized into 128x128 using an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://www.kaggle.com/gpiosenka/100-bird-species.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has not been used yet. However, the original dataset that we used to build this one has already been used for birds classification.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC0 Public Domain (https://creativecommons.org/publicdomain/zero/1.0/). Further information about licenses can be found in the "info.json" meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for LR_AM.DOG Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Dustin Carrión created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of dogs on natural backgrounds. In the original dataset, the resolution of the instances varies from 100x105 to 2448x3264, with an average resolution of 385x442.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 4800 instances from 120 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image in the formatted dataset. The instances were preprocessed, i.e., cropped and resized into 128x128 with a smart cropping algorithm and an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance has a category/label, which is provided with the meta-data. A category corresponds to a dog breed.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are some duplicated images in the original dataset, but they were removed in the formatted dataset.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

In the original dataset, the images and bounding boxes were downloaded from ImageNet.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original authors of the Stanford Dogs dataset: Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, Li Fei-Fei.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The exact date is unknown but all the data were collected before before June 2011.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 using smart cropping and anti-aliasing filter. Moreover, in the original dataset, each image was examined to confirm whether or not it matched images from Wikipedia and shared similar features to the other images in the same category. Degenerate or unusual images (distorted colors, very blurry or noisy, largely occluded, extreme close-ups) were removed manually. All duplicated images, within and between categories, were removed. The bounding boxes on ImageNet were annotated and verified through Amazon Mechanical Turk.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data corresponds to the original dataset (http://vision.stanford.edu/aditya86/ImageNetDogs/), while the preprocessed data correspond to the formatted dataset.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

This dataset has not been used yet. However, the original dataset from which we sampled it has already been used for fine-grained image categorization.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and for traditional dogs classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original dataset is cited and is released under the license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the "info.json" meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for LR_AM.AWA Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Dustin Carrión created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of animals on natural backgrounds. In the original dataset, the resolution of the instances varies from 100x100 to 1893x1920, with an average resolution of 660x754.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of all instances from 50 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The instances were preprocessed, i.e., cropped and resized into 128x128 with a smart cropping algorithm and an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance has a category/label, which is provided with the meta-data. A category corresponds to an animal.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

In the original dataset, the data (images of animals) were collected in 2016 from public web-sources (Flickr, Wikimedia, ...), considering only images with a license that allows free use and redistribution.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

N/A.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original data were collected in 2016, but there is no specification of the exact period.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The images are cropped and resized into 128x128 using a smart cropping algorithm and an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source `https://cvml.ist.ac.at/AwA2/`.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

This dataset has not been used yet. The original from which we sampled the current dataset has already been used for zero-shot classification.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and for traditional animal classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : Creative Commons (https://cvml.ist.ac.at/AwA2/). Further information about licenses can be found in the "info.json" meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for SM_AM.PLK dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research [1], grant number 20HR0134, Labex Digicosme [2] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn [3] a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of planktons.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 102 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images per class. Mini version consists of all classes that have at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[1] https://anr.fr/
[2] https://digicosme.cnrs.fr/
[3] http://www.chalearn.org/

The dataset is sampled from a large plankton dataset [4] which consists of 3.5 million images. The extracted dataset is representative in terms of classes, i.e. it represents all classes in original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128×128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a label which is provided with the images in meta-data. The label of each plankton is provided by researchers at the Woods Hole Oceanographic Institution[5].

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[4] https://github.com/hsosik/WHOI-Plankton
[5] https://www.whoi.edu/

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of plankton) is collected by researchers at the Woods Hole Oceanographic Institution[6]. Imaging FlowCytobot (IFCB) was used for the data collection. Complete process and mechanism is described here [10].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Researchers at the Woods Hole Oceanographic Institution[6]: Heidi M. Sosik, Emily E. Peacock and Emily F. Brownlee were involved in the data collection process.

[6] https://www.whoi.edu/

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected between 2006 and 2014.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e matching backgrounds are added to portrait and landscape images to make squared images and then the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://github.com/hsosik/WHOI-Plankton

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated Github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of plankton.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : MIT License (https://github.com/hsosik/WHOI-Plankton/blob/master/LICENSE). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on Github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the Github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for SM_AM.INS_2 dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB pest insects.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 102 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128x128 RGB image. The instances are preprocessed: resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the insect species.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data were acquired from the Internet, only the top-2,000 results were kept. Other images were collected from agriculture, insect science websites and videos. Then 6 volunteers manually filtered the candidate images, they deleted images containing none or more than one insect per image and also corrupted images. Finally, the images were annotated by 8 agricultural experts, only the images annotated with the same label by at least 5 of the experts are kept.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Volunteers for the first selection of images, 8 experts for the final annotation. Probably the original authors of the benchmark : Xiaoping Wu, Chi Zhan, Yu-Kun Lai, Ming-Ming Cheng, Jufeng Yang.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The paper was released in 2019.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed. The images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://github.com/xpwu95/IP102

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated Github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of insects.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for academic research as long as the original paper is cited and we redistribute it under license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on Github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the Github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for SM_AM.INS dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[7], grant number 20HR0134, Labex Digicosme[8] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn[9] a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of insects on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 114 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is sampled from a large insects dataset which consists of 290,000 images. The extracted dataset is representative in terms of classes, i.e. it represents all classes in original dataset.

[7] https://anr.fr/
[8] https://digicosme.cnrs.fr/
[9] http://www.chalearn.org/

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128×128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the insect's specie, family or super-family. Along with the category, the meta-data also has super-category which describes the super-class of the insect which can be a family, super-family or Order.
The categories and super-categories are assigned by citizen scientists and entomologists of the *SPIPOL science project*[10], a Citizen Science program created and managed by the *French Natural History National Museum*[11] with the help of a entomologist NGO. The process of assignation goes by multiple assignation, versioning of assignations and possibility to restart from scratch assignation at any time if doubtful. More information and original data is accessible on the SPIPOL Website[10].

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[10] https://www.spipoll.org/
[11] https://www.mnhn.fr/fr

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of insects) is collected by citizen scientists. They are asked to take pictures of insects and other bugs on flowers and to upload these pictures with a proposal for assignation. Then, pictures are assigned by pairs and external experts. 3 same assignations are considered to be valid but at anytime, anyone can propose a new assignation. The pictures are uploaded to the hosting server by using the website[10] or android application[12] or iphone application[13].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

[12] https://play.google.com/store/apps/details?id=fr.eneo.spipoll
[13] https://apps.apple.com/fr/app/spipoll/id1495843067

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Volunteers and Spipoll team[10] were involved in the data collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected between May 2010 and April 2019.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed: the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://www.spipoll.org

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of insects.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC BY NC 2.0 (https://www.spipoll.org/mentions-legales). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on Github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors

of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the Github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT.FLW dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Felix Mohr created the dataset. At the time of creation, Felix Mohr is associated professor at Universidad de La Sabana, Chía, Colombia.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[14], grant number 20HR0134, Labex Digicosme[15] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn[16] a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of different flowers.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 102 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

[14] https://anr.fr/
[15] https://digicosme.cnrs.fr/
[16] http://www.chalearn.org/

Each instance is a 128×128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a label/category which is provided with the images in meta-data. The label of each flower is provided by the original creators of the flowers dataset[17].

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

[17] https://www.robots.ox.ac.uk/~vgg/data/flowers/

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of flowers) is collected by the original creators of the Flowers dataset[18]. The images are collected from various websites and some are photographed.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Authors of the paper 'Automated Flower Classification over a Large Number of Classes'[7]: M. Nilsback and A. Zisserman were involved in the data collection process. '

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected before 2008.

[18] https://www.robots.ox.ac.uk/~vgg/data/flowers/

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset, however it can be accessed from its original source https://www.robots.ox.ac.uk/~vgg/data/flowers/102/.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, these datasets can be used for classification tasks of flowers.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license: GNU General Public License Version 2 (https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website (https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT.PLT_NET Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Felix Herron created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302)

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of plants on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 1081 classes, here referred to as scientific names. There are a variable number of images per class, ranging from 4 to 9011. This dataset contains the most represented 25 classes. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances were preprocessed i.e. resized from their initial (mainly-square) sizes into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a scientific name (category) which is provided with the images as meta-data.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There is no explicit link of the genus of the plants; this could be added using external sources.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The dataset was constructed based on data submitted by users around the world to the Pl@ntNet Project (`https://plantnet.org/en/`) by over 2,000,000 citizen botanists as well as experts, who furnish images and proposed labels. The annotations are supported by a system of volunteers from over 170 countries who reinforce each other's expertise using weighted reliability scores: anyone can upload an image, others must indicate that they agree with its classification, and the expertise of the annotator determines the weight of the vote given to a label. There are 2.03 annotators per image for the Pl@ntNet Project as a whole.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Contributors to Pl@ntNet are volunteers; however, Pl@ntNet does retain a staff of full-time engineers and researchers. It does not appear that they are directly involved in the curation of the dataset, however.

The dataset was compiled by Camille Garcin, alexis joly, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, Joseph Salmon; all scientists at the University of Montpellier (among other institutions), as well as Pierre Bonnet, of CIRAD and

AMAP.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected between 2009 and 2021; it is not clear whether the effort was temporally consistent.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://gitlab.inria.fr/cgarcin/plantnet_dataset

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Our formatted dataset was not used for any tasks yet. However, Pl@ntNet Project data has been cited hundreds of times since the project's inception, in various botanical and image classification projects, such as by Affouard et. al in "Pl@ntNet app in the era of deep learning" or by Heredia in "Large-Scale Plant Classification with Deep Neural Networks". The paper publishing this particular dataset has no apparent citations, however.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

This dataset is meant to be used as a reinforcement for any image classification tasks for which ImageNet might otherwise be used; hence, classification is its primary focus. Perhaps generative algorithms could be trained on this dataset as well.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license: Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/legalcode). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors

of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT.FNG Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Felix Herron created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of fungi on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 182 classes. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances were preprocessed i.e. resized from their initial (variable) sizes into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a scientific name (category) as well as a genus (super category), as well as other metadata which was omitted for this project, which is provided with the images in meta-data.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All images classified as the same species are also classified as the same genus.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There are several omitted metadata fields which are redundant but they are irrelevant for this project.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The dataset was constructed based on submissions to the Atlas of Danish Fungi (https://svampe.databasen.org/), an organization "supported by more than 3,300 volunteers" who furnish images and proposed labels. The annotations are supported by a system of volunteers who reinforce each other's expertise using reliability scores: anyone can upload an image, others must indicate that they agree with its classification. Once consensus has been reached, a group of experts signs off on most classifications.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Contributors to the Atlas are volunteers; no evidence of compensation for experts is indicated either.

The dataset itself was compiled by Lukáš Picek, Lukáš Picek, of the University of West Bohemia; Milan Šulc and Jiří Matas, of the CTU in Prague; Thomas S. Jeppesen of the GBIF; Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev of the University of Copenhagen.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected between 1963 and 2020, although the vast majority (99.4%) of images were taken between 2009 and 2020.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with anti-aliasing filter. Furthermore, as previously described, all images from species with fewer than 40 instances were

removed, and only the top-25 classes were retained.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://sites.google.com/view/danish-fungi-dataset

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

This dataset is meant to be used in heavily imbalanced machine learning classification settings. However, it could be used from a botanical, rather than purely ML perspective, as an analysis as to which plants are most easily recognizable.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : BSD-3-Clause License (https://github.com/picekl/DanishFungiDataset/blob/main/LICENSE). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT_DIS.PLT_VIL dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Phan Anh Vu created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[19], grant number 20HR0134, Labex Digicosme[20] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302) and ChaLearn[21] a 501(c)(3) non-for-profit California organization.

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are RGB images of leaves.

How many instances are there in total (of each type, if appropriate)?

There are 37 categories in total. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[19] https://anr.fr/
[20] https://digicosme.cnrs.fr/
[21] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a RGB image.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No recommendation for data split.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Some leaves may appear in several photos. Backround and lighting condition may have some effects.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No sensitive information.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image captured by camera.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Leaves are removed from the plant, then placed on a paper sheet with gray background color. Photos are taken outside, under full light, with a camera.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Members of the Plant Village project [22] collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original collection Plant Village is released in 2015.

[22] https://data.mendeley.com/datasets/tywbtsjrjv/1

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, each original image was converted to a 128x128 pixel RGB image, by cropping it around the region of interest in a square, and reducing its dimension with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://data.mendeley.com/datasets/tywbtsjrjv/1

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and Leaf disease classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC0 1.0 (https://creativecommons.org/publicdomain/zero/1.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT_DIS.MED_LF dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Phan Anh Vu created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[23], grant number 20HR0134, Labex Digicosme[24] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302) and ChaLearn[25] a 501(c)(3) non-for-profit California organization.

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are RGB images of leaves.

How many instances are there in total (of each type, if appropriate)?

There are 27 categories in total. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[23] https://anr.fr/
[24] https://digicosme.cnrs.fr/
[25] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a RGB image.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

No recommendation for data split.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No sensitive information.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image captured by mobile phone camera.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The leaves are plucked from different plants of the same species, then placed on white uniform background.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Authors of the Medicinal Leaf dataset [26] collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original collection Medicinal Leaf is released in 2020.

[26] https://data.mendeley.com/datasets/nnytj2v3n5/1

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, each original image was converted to a 128x128 pixel RGB image, by cropping it around the region of interest in a square, and reducing its dimension with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://data.mendeley.com/datasets/nnytj2v3n5/1

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and leaf diseases classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for PLT_DIS.PLT_DOC Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

María Belén Guaranda created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of healthy and damaged leaves on natural backgrounds or scraped from the Internet .

How many instances are there in total (of each type, if appropriate)?

The dataset consists of a total of 27 classes/categories. There are at least 54 instances/images per class, and a maximum of 188. Total count of instances is 2,579. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter, using the PIL library in Python (Lanczos algorithm).

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided in a CSV file. The category describes the plant's species or the name of the disease. There are no super categories.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

For the images scraped from the internet, it would have been useful to have some metadata or field that indicates the license or source of the image, in order to not violate any copyright and identify those images that are not taken in a natural way (e.g.: with no prepared set up, with the sun as the source of light). This information was not available in the original dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

But the original dataset is provided in a defined split for the train and test sets and we saved this information on the csv file associated with the dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Noise: some images have a watermark. There are also images of a very low resolution, so the leaf in the image may appear very blurry and with no defined shape.

Errors: the authors of the original dataset point out that some of the images could be wrongly labeled, as they performed the labeling task, but they are not experts on the field.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of plants) were collected by citizen scientists from Google Images and Ecosia. 20,900 images were collected by using scientific and common names of 38 classes mentioned in the dataset by Mohanty et al. [9]. Four users filtered the images by selecting images based on their metadata on the website and guidelines mentioned on APSNet. Using referred APS' prior literature, they accordingly classified images. Some of the most important factors for classification, mentioned by the authors, were the color, area and density of the diseased part and shape of the species. Every image was then checked by two individuals according to the guidelines to reduce labeling errors. Classes with less than 50 images were discarded. Afterwards, on each picture, using a library called LabelImg tool, they manually draw bounding boxes to identify each single instance of a leaf. The filename of each picture, the coordinates of the bounding boxes of each leaf and the class each leaf belongs to, were provided in a CSV file, along with the original pictures, to build the dataset of

leaves.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Scientists were involved in the data collection.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were collected between May 2010 and April 2019.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://github.com/pratikkayal/PlantDoc-Object-Detection-Dataset

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The formatted version of the dataset was not used for any task. However, the original dataset we used to create this formatted version was used for leaves detection and plant diseases classification.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of plant diseases.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : Creative Commons Attribution 4.0 International (https://github.com/pratikkayal/PlantDoc-Object-Detection-Dataset/blob/master/LICENSE.txt). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for MCR.BCT Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Romain Mussard created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of colony bacteria obtained by microscopy.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of a total 33 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/class. The Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. The Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. splitting of the original images with a patch size of 650x650 pixels, resized into 128x128 with anti-aliasing filter, background deletion and stain normalization.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the bacteria species.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022. But the data should respect a specific split to avoid bias because some images derive from the same original (mother) image. We proposed a split in the csv file. The mother image of each image is represented by the first number of the file name.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The original data were collected by the Chair of Microbiology of the Jagiellonian University in Krakow, Poland. Stained using the Gramm's method. The images were taken with Olympus CX31 Upright Biological Microscope equipped with a SC30 camera (Olympus Corporation, Japan). They were evaluated using a 100 times objective under oil-immersion (Nikon50, Japan).

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data were collected by Bartosz Zieliński, Anna Plichta, Krzysztof Misztal, Przemysław Spurek, Monika Brzychczy-Włoch, Dorota Ochońska

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were first used in an article on May 27, 2017 (We do not have additional information about the data collection procedure).

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

N/A

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are split in multiple sub-images, we then cleaned the data by removing blank images with no Region Of Interest (using otsu's method [8] to binarize the image and separate background and bacteria).

Each of the obtained images are then resized with anti-aliasing techniques. We remove the background and replace it by a white background and then use stain normalization techniques.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://www.aftermarket.pl/domena/misztal.edu.pl/?_track=b465acbe1e83b8a89ff9c1d238cc99d6 or https://github.com/gallardorafael/DIBaS-Dataset

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and for bacteria colony classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for MCR.PRT Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Felix Mohr created the dataset. At the time of creation, Felix Mohr is associated professor at Universidad de La Sabana, Chía, Colombia.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB microscopy images showing subcellular structures of human body proteins.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 21 classes/categories. Total count of instances is 15,050 images. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is an adjusted sub-dataset with instances from the proteinatlas.org project. From the original dataset, which is a multi-label dataset, all instances with more than one label were removed and all instances of classes that do not sum up at least 40 labeled instances. The original data has

higher resolutions, so the images were resized to match into the benchmark format.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the type of cell structure visible in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is a microscopy image and is hence directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The description here is a quote from the original paper Thul et al. (A subcellular map of the human proteome), 2017

Fluorescent images were acquired with a Leica SP5 confocal microscope (DM6000CS) equipped with a 63x HCX PL APO 1.40 oil CS objective (Leica Microsystems, Mannheim, Germany). The settings for each image were as follows: Pinhole 1 Airy unit, 16-bit acquisition, and a pixel size of 0.08 micro-m. The detector gain measuring the signal of each antibody was adjusted to a maximum of 800 V to avoid strong background noise. The majority of the images were acquired manually from at least two representative field-of-views (FOVs). For proteins displaying single cell variations in their expression pattern, at least six different FOVs were acquired. A small part of the plates were imaged automatically using the MatrixScreener M3 in LAS AF software (Leica Microsystem, Mannheim, Germany). Here, z-stacks at six FOVs were acquired and afterward two images were manually selected for display in the Cell Atlas. All images on the Cell Atlas are unprocessed with a small compression due to conversion from TIFF to JPEG file format.

The subcellular location of each protein was manually determined based on the signal pattern and relation to the markers for nucleus (DAPI), microtubules, and endoplasmic reticulum. The annotated

locations were as follows: actin filaments, aggresome, cell junctions, centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, endoplasmic reticulum, focal adhesions, Golgi apparatus, intermediate filaments, lipid droplets, microtubule organizing center (MTOC), microtubules, microtubule ends, midbody, midbody ring, mitochondria, mitotic spindle, nuclear bodies, nuclear membrane, nuclear speckles, nucleolar fibrillar center, nucleolar rim, nucleoli, nucleoplasm, nucleus, plasma membrane, rods and rings, and vesicles. If more than one location was detected, they were defined as main or additional location depending on the relative signal strength between the location and the most common location when including all cell lines. Variation between single cells were annotated either as a variation in the intensity or spatial distribution based on a visual inspection. The staining was not annotated if considered negative or unspecific.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Deterministic sub-sample, which includes all instances that satisfy the single-label condition and belong to a class with at least 40 instances.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original author of the subcellular dataset : Peter Thul and the co-authors of the above papers (over 30 authors in total).

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were released in May 2017.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://console.cloud.google.com/storage/browser/kaggle-human-protein-atlas

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, the github repository (https://meta-album.github.io/) will be active once the dataset is publicly released, it will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and classification tasks of subcellular structures.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future

user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Github repository (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the github repository (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and we redistribute it under license : CC BY-SA 3.0 (https://www.proteinatlas.org/about/licence). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository (https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding github repository (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available at the github repository (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the github repository (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for MCR.PNU Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Romain Mussard created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of Human Tissue obtained by microscopy.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of a total 19 classes/categories for a total amount of 7753 images. There are at least 101 images by classes. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter, and preprocessed using stain normalization techniques.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the bacteria species.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The original data were collected by a group of searchers from different laboratories. The data is sampled over an aggregated set of publicly available nucleus classification and detection datasets.

Initially the dataset was design for nuclei segmentation and classification

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The images are patches of a selection of WSI representing 19 different tissue types.
Initially 2,000 visual fields were sampled from more than 20,000 WSIs

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data were collected by Jevgenij Gamper , Navid Alemi Koohbanani , Ksenija Benes, Simon Graham, Mostafa Jahanifar, Seyyed Ali Khurram, Ayesha Azam, Katherine Hewitt and Nasir Rajpoot

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

First use in an article on July 2019

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Other sources : websites

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images were resized with anti-aliasing techniques and stain normalized.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://jgamper.github.io/PanNukeDataset/

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and for human tissu classification. Furthermore, the original datset was desgin as a nuceli segmentation dataset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : Attribution-NonCommercial-ShareAlike 4.0 International (https://creativecommons.org/licenses/by-nc-sa/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for REM_SEN.RESISC dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Phan Anh Vu created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[27], grant number 20HR0134, Labex Digicosme[28] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302) and ChaLearn[29] a 501(c)(3) non-for-profit California organization.

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are RGB remote sensing images.

How many instances are there in total (of each type, if appropriate)?

There are 45 categories in total. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[27] https://anr.fr/
[28] https://digicosme.cnrs.fr/
[29] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a RGB image.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Adjacent areas may appear in several patches.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No sensitive information.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an aerial photo of scene and landscape.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The images are retrieved from Google Earth.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Members of the RESISC45 project [30] collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original collection RESISC45 is released in 2017.

[30] https://https://gcheng-nwpu.github.io

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://gcheng-nwpu.github.io/

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and Aerial image classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for REM_SEN.RSICB dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Phan Anh Vu created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[31], grant number 20HR0134, Labex Digicosme[32] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302) and ChaLearn[33] a 501(c)(3) non-for-profit California organization.

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are RGB remote sensing images.

How many instances are there in total (of each type, if appropriate)?

There are 45 categories in total. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[31] https://anr.fr/
[32] https://digicosme.cnrs.fr/
[33] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a RGB image.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Adjacent areas may appear in several patches.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No sensitive information.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an aerial photo of scene and landscape.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The images are retrieved from Google Earth and Bing Map.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Members of the RSICB project [34] collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original collection RSICB128 is released in 2020.

[34] https://github.com/lehaifeng/RSI-CB

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://github.com/lehaifeng/RSI-CB

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and Aerial image classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we redistribute it under license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for REM_SEN.RSD dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Phan Anh Vu created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[35], grant number 20HR0134, Labex Digicosme[36] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302) and ChaLearn[37] a 501(c)(3) non-for-profit California organization.

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are RGB remote sensing images.

How many instances are there in total (of each type, if appropriate)?

There are 46 categories in total. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[35] https://anr.fr/
[36] https://digicosme.cnrs.fr/
[37] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a RGB image.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Adjacent areas may appear in several patches.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

N/A

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

N/A

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No sensitive information.

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an aerial photo of scene and landscape.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The images are retrieved from Google Earth and Tianditu.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Members of the RSD46 project [38] collected the data.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original collection RSD46 is released in 2017.

[38] https://github.com/RSIA-LIESMARS-WHU/RSD46-WHU

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

No

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://github.com/RSIA-LIESMARS-WHU/RSD46-WHU

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and Aerial image classification.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we redistribute it under license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for VCL.CRS Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Philip Boser created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB cars images acquired from Flickr, Google, and Bing.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 196 classes/categories. Total count of instances is 7,840 images. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the car model.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data were acquired from Flickr, Google, and Bing. All the sample images are labeled by non specialists but all of them were selected thanks to qualification tasks.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original authors of the car dataset : Jonathan Krause , Michael Stark, Jia Deng , and Li Fei-Fei.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were released in May 2013.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source `http://ai.stanford.edu/~jkrause/cars/car_dataset.html`.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and classification tasks of cars.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : ImageNet License (https://ai.stanford.edu/~jkrause/cars/car_dataset.html). Further information about licenses can be found in the "info.json" meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for VCL.APL Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Philip Boser created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, by the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB remote sensing airplanes images acquired from Google Earth satellite imagery.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 21 classes/categories. The total count of instances is 840 images. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/classes. The mini version consists of all classes that have at least 40 images per class and randomly selected 40 images per class. The Micro version consists of 20 randomly selected classes from the Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the airplane model.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise, or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However, it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data were acquired from Google Earth satellite imagery and manually expanded. All the sample images are carefully labeled by seven specialists in the field of remote sensing image interpretation. Each image contains one and only one complete aircraft.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

Seven specialists in the field of remote sensing images interpretation and probably the original author of the benchmark data set for aircraft type recognition from remote sensing images : Zhi-Ze Wu, Shou-Hong Wan, Xiao-Feng Wang, Ming Tan, Le Zou, Xin-Lu Li, Yan Chen

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., a recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were released in May 2019.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source https://zenodo.org/record/3464319.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, these datasets can be used for continual learning, meta-learning, cross-domain meta-learning, and classification tasks of airplanes.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that it is appropriate for conducting scientific research other than research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/legalcode). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case, updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided a complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of newly constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for VCL.BTS Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Dustin Carrión created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB boats images acquired from Flickr, Google, and Bing.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 26 classes/categories. Total count of instances is 1,040 images. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the boat model.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, the dataset has no errors, noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data were sampled from 2 million images uploaded by people on a community website. A semi-automatic clustering scheme was used to combine the original 109 classes into 26 super-classes.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original author of this dataset : Erhan Gundogdu, Berkan Solmaz, Veysel Yücesoy & Aykut Koç

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data were released in May 2016.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source `https://github.com/avaapm/marveldataset2016`.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for continual learning, meta-learning, cross-domain meta-learning and classification tasks of boats.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we released it under license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on our github repository (details provided on Meta Album website https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for MNF.TEX Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[39], grant number 20HR0134, Labex Digicosme[40] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn[41] a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of textures.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 64 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/class. The mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. The micro version consists of 20 randomly selected classes from the Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[39] https://anr.fr/
[40] https://digicosme.cnrs.fr/
[41] http://www.chalearn.org/

This dataset is a combination of 4 textures datasets : The dataset is sampled from 4 texture datasets: KTH-TIPS [3], KTH-TIPS 2 [6], Kylberg Textures Dataset [4] and UIUC Tectures Dataset [5].

The dataset contains all possible instances of the original datasets.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an 128×128 RGB image. The instances are preprocessed, i.e., resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a label which is provided with the images in meta-data.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data in all four datasets is collected in laboratory conditions, i.e., images were captured in controlled environment with configurable brightness, luminosity, scale and angle.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The *KTH-TIPS* dataset was collected by Mario Fritz and *KTH-TIPS 2* dataset was collected by P. Mallikarjuna and Alireza Tavakoli Targhi. Both of these datasets were prepared under the supervision of Eric Hayman and Barbara Caputo. The data for *Kylberg Textures Dataset* and *UIUC Tectures*

Dataset data were collected by the original authors of these datasets.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

KTH-TIPS and *KTH-TIPS 2* datasets were created in 2004 and 2006 respectively. *Kylberg Textures Dataset* was created in September 2010 and *UIUC Tectures Dataset* in August 2005.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

N/A

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing

of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed, i.e, the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset, however it can be accessed from its original source: https://github.com/abin24/Textures-Dataset

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?
N/A

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of textures.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?
N/A

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we redistribute it under the CC-BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?
N/A

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of the Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in the Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via the following email address: meta-album@chalearn.org.

Any other comments?

N/A

Datasheet for MNF.TEX_DTD dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[42], grant number 20HR0134, Labex Digicosme[43] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn[44] a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of textures.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 64 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/class. The mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. The micro version consists of 20 randomly selected classes from the Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[42] https://anr.fr/
[43] https://digicosme.cnrs.fr/
[44] http://www.chalearn.org/

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an 128×128 RGB image. The instances are preprocessed, i.e., resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a label which is provided with the images in meta-data.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of textures) is collected from Google and Flicker. The data were annotated using Amazon Mechanical Turk [2].

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data were collected by the original authors of the paper "Describing Textures in the Wild"[2]. The data collection process is mentioned on the dataset overview page

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data collection process started in June and July 2012.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

N/A

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed, i.e, the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset, however it can be accessed from its original source: https://www.robots.ox.ac.uk/~vgg/data/dtd/

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?
N/A

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of textures.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?
N/A

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we redistribute it under the CC-BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?
N/A

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of the Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in the Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via the following email address: meta-album@chalearn.org.

Any other comments?

N/A

Datasheet for MNF.TEX_ALOT dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Ihsan Ullah created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research[45], grant number 20HR0134, Labex Digicosme[46] project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn[47] a 501(c)(3) non-for-profit California organization.

Any other comments?

N/A

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of textures.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 250 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/class. The mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. The micro version consists of 20 randomly selected classes from the Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

[45] https://anr.fr/
[46] https://digicosme.cnrs.fr/
[47] http://www.chalearn.org/

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an 128×128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a label which is provided with the images in meta-data.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in labels.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

N/A

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data (images of textures) is collected in laboratory conditions, i.e., images were captured in controlled environment with configurable brightness, luminosity, scale and angle.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data were collected by the original creators of the Texture-ALOT dataset [1]. The data collection process is mentioned on the dataset overview page(`https://aloi.science.uva.nl/public_alot/`).

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The period of data collection is unknown.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

N/A

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed, i.e, the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset, however it can be accessed from its original source: `https://aloi.science.uva.nl/public_alot/`

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?
N/A

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of textures.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?
N/A

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research as long as the original paper is cited and we redistribute it under the CC-BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?
N/A

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of the Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in the Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via the following email address: meta-album@chalearn.org.

Any other comments?

N/A

Datasheet for HUM_ACT.SPT Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Jilin He created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of human sports on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of total 73 classes/categories. There are 3 versions of this dataset in Meta-Album. Extended version consists of all classes and all images/class. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category represents a kind of sport.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset contains images of people doing sports.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Images were gathered from internet searches. The images were scanned with a duplicate image detector program written by the author(https://www.kaggle.com/gpiosenka/sports-classification). All images were then resized to 224X224X3 and converted to jpg format.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Gerry(https://www.kaggle.com/gpiosenka) was involved in the data collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

N/A

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, but only very coarsly grained. No people are identifiable.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The were originally obtained from websites.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The images are from publicly broadcasted sports events.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e the images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source. https://www.kaggle.com/datasets/gpiosenka/sports-classification

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of human actions/sports.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC0 1.0 Public Domain (https://creativecommons.org/publicdomain/zero/1.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for HUM_ACT.ACT_40 Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Jilin He created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, by the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of human actions on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of a total 39 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/classes. Mini version consists of all classes that has at least 40 images per class and randomly selected 40 images per class. Micro version consists of 20 randomly selected classes from Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category represents a specific daily human action.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise, or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However, it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset consists of images of 39 kinds of human actions.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes, some images are of celebrities that could be identified.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

N/A

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The original dataset is associated with the paper : B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unsure, but crawled from websites before 2011.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data is gathered from web crawls.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e

1. The images are cropped by using human posture detection algorithms and specific cropping rules to make sure the shape of the image is square and contains the target action in it;
2. 1 class is removed from the original dataset due to insufficient quantities after step 1;
3. For Mini version 40 images are randomly chosen from each class; For Micro version 20 classes are randomly chosen from Mini version.
4. The images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source http://vision.stanford.edu/Datasets/40actions.html

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Not in this version, but it is derived from the Stanford 40 Actions datasets, which are used in several scientific works on action recognition.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, these datasets can be used for continual learning, meta-learning, cross-domain meta-learning, and classification tasks of human actions.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that it is appropriate for conducting scientific research other than research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (to benchmark machine learning models).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research, citation of the original paper is mandatory, and we release it under license : CC-BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on the Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case, updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided a complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of newly constructed datasets using the defined protocol are given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for HUM_ACT.ACT_410 Dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Jilin He created the dataset, under the supervision of Professor Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, by the TAU team, as part of the HUMANIA project, funded by the French research agency ANR and Labex Digicosme. ChaLearn also supported the creation of the dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research (https://anr.fr/), grant number 20HR0134, Labex Digicosme (https://digicosme.cnrs.fr/) project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex ParisSaclay (ANR11IDEX000302) and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128x128 RGB images of human sports on natural backgrounds.

How many instances are there in total (of each type, if appropriate)?

The dataset consists of a total 29 classes/categories. There are 3 versions of this dataset in Meta-Album. The extended version consists of all classes and all images/classes. The mini version consists of all classes that have at least 40 images per class and randomly selected 40 images per class. The Micro version consists of 20 randomly selected classes from the Mini version with 40 images per class.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains all possible instances of the original dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is 128x128 RGB image. The instances are preprocessed i.e. resized into 128x128 with an anti-aliasing filter.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category represents a kind of sport.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all information is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No, there are no suspected errors, sources of noise, or redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However, it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset contains pictures of people doing sports.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Each image was extracted from a YouTube video and provided with preceding and following unannotated frames.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele team were involved in the data collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

N/A

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, the data is preprocessed i.e

1. The images are cropped by specific cropping rules to make sure the shape of the image is square and contains the target action in it;
2. For the Mini version 40 images are randomly chosen from each class; For the Micro version, 20 classes are randomly chosen from the Mini version.
3. The images are resized into 128x128 with an anti-aliasing filter.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data is not released with the preprocessed dataset however it can be accessed from its original source http://human-pose.mpi-inf.mpg.de

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository. Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, these datasets can be used for classification tasks of human actions/sports.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that it is appropriate for conducting scientific research other than research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : Simplified BSD License (http://human-pose.mpi-inf.mpg.de/bsd.txt). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on the Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case, updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided a complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of newly constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for OCR.MD_MIX datasets

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Haozhe Sun created the dataset, under the supervision of Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR. ChaLearn also supported the development of the software.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research, https://anr.fr/), grant number 20HR0134 and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of synthetic printed characters.

How many instances are there in total (of each type, if appropriate)?

OmniPrint-MD-mix has 28240 images from 706 classes (each class has 40 images).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

These datasets are synthesized from the data synthesizer OmniPrint, thus they can be viewed as a sample of instances from all the possible images given the nuisance parameters (fonts, styles, noises, etc.). These datasets are representative of such images because the synthesis parameters of each instance were uniformly sampled, no further selection was performed.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128×128 RGB image. Each image contains one single character from a certain script, rendered in a particular way (background, foreground, distortions, noises).

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the character identity. Along with the category, the meta-data also has super-category (alphabet or subset of alphabet) and various synthesis parameters.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all of the metadata is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories, all provided.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

We intentionally introduced various transformations and noises to each image instance. The transformation parameter space is large so there is little chance that two instances are identical.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is synthesized by OmniPrint. Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data is synthesized using the data synthesizer OmniPrint. The involved Unicode characters were manually selected from the Unicode standard, which constitutes a set of characters from several languages around the world. The involved fonts were downloaded from a manually-defined list of URLs, the downloaded fonts were then filtered by a python program in order to filter corrupted fonts. Several distortions and noises were involved, including affine and perspective transformations, random elastic transformations, natural background, foreground text filling, etc.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data is synthesized by a data synthesizer OmniPrint. The sampling is uniformly random in the given transformation parameter space.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data is synthesized by a computer software. However the design and implementation of the software, the choice of characters and fonts involve the authors of OmniPrint [11].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The datasets were synthesized on June 24, 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing/cleaning/labeling was performed. The datasets are made available as they were synthesized. No feature extraction or removal of instances was done.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository and OmniPrint repository (https://github.com/SunHaozhe/OmniPrint). Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of a large number of characters, and for domain adaptation tasks. Furthermore, as the meta-data can serve as labels, other kinds of classification or regression problems can also be considered e.g. classification of fonts, classification of languages, regression of rotation angle, regression of horizontal shear, etc. Finally, the datasets can be used to study disentangling the label (class character) from the nuisance variables (font, style, distortions).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC BY 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for OCR.MD_5_BIS datasets

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Haozhe Sun created the dataset, under the supervision of Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR. ChaLearn also supported the development of the software.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research, https://anr.fr/), grant number 20HR0134 and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of synthetic printed characters.

How many instances are there in total (of each type, if appropriate)?

OmniPrint-MD-5-bis has 28240 images from 706 classes (each class has 40 images).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Theis datasets is synthesized from the data synthesizer OmniPrint, thus they can be viewed as a sample of instances from all the possible images given the nuisance parameters (fonts, styles, noises, etc.). These datasets are representative of such images because the synthesis parameters of each instance were uniformly sampled, no further selection was performed.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128×128 RGB image. Each image contains one single character from a certain script, rendered in a particular way (background, foreground, distortions, noises).

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the character identity. Along with the category, the meta-data also has super-category (alphabet or subset of alphabet) and various synthesis parameters.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all of the metadata is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories, all provided.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

We intentionally introduced various transformations and noises to each image instance. The transformation parameter space is large so there is little chance that two instances are identical.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is synthesized by OmniPrint. Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data is synthesized using the data synthesizer OmniPrint. The involved Unicode characters were manually selected from the Unicode standard, which constitutes a set of characters from several languages around the world. The involved fonts were downloaded from a manually-defined list of URLs, the downloaded fonts were then filtered by a python program in order to filter corrupted fonts. Several distortions and noises were involved, including affine and perspective transformations, random elastic transformations, natural background, foreground text filling, etc.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data is synthesized by a data synthesizer OmniPrint. The sampling is uniformly random in the given transformation parameter space.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data is synthesized by a computer software. However the design and implementation of the software, the choice of characters and fonts involve the authors of OmniPrint [11].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The datasets were synthesized on June 24, 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing/cleaning/labeling was performed. The datasets are made available as they were synthesized. No feature extraction or removal of instances was done.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository and OmniPrint repository (https://github.com/SunHaozhe/OmniPrint). Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of a large number of characters, and for domain adaptation tasks. Furthermore, as the meta-data can serve as labels, other kinds of classification or regression problems can also be considered e.g. classification of fonts, classification of languages, regression of rotation angle, regression of horizontal shear, etc. Finally, the datasets can be used to study disentangling the label (class character) from the nuisance variables (font, style, distortions).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC BY 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data files. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

Datasheet for OCR.MD_6 dataset

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for competition and benchmark purposes as part of the Meta-Album meta-dataset. The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and conduct benchmarks, particularly in: few-shot learning, meta-learning, continual learning, transfer learning, and image classification.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Haozhe Sun created the dataset, under the supervision of Isabelle Guyon. The work was performed at LISN laboratory, Université Paris-Saclay, France, in the TAU team, as part of the HUMANIA project, funded by the French research agency ANR. ChaLearn also supported the development of the software.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANR (Agence Nationale de la Recherche, National Agency for Research, https://anr.fr/), grant number 20HR0134 and ChaLearn (http://www.chalearn.org/) a 501(c)(3) non-for-profit California organization.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances are 128×128 RGB images of synthetic printed characters.

How many instances are there in total (of each type, if appropriate)?

OmniPrint-MD-6 has 28120 images from 703 classes (each class has 40 images).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

These datasets are synthesized from the data synthesizer OmniPrint, thus they can be viewed as a sample of instances from all the possible images given the nuisance parameters (fonts, styles, noises, etc.). These datasets are representative of such images because the synthesis parameters of each instance were uniformly sampled, no further selection was performed.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a 128×128 RGB image. Each image contains one single character from a certain script, rendered in a particular way (background, foreground, distortions, noises).

Is there a label or target associated with each instance? If so, please provide a description.

Yes, each instance has a category/label which is provided with the images in meta-data. The category describes the character identity. Along with the category, the meta-data also has super-category (alphabet or subset of alphabet) and various synthesis parameters.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all of the metadata is provided for each instance.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All relationships are contained in categories and super-categories, all provided.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The data has no splits because the splits are generated on the fly in NeurIPS Cross-Domain Meta-Learning Competition 2022.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

We intentionally introduced various transformations and noises to each image instance. The transformation parameter space is large so there is little chance that two instances are identical.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. It will exist, and remain constant, over time once we release it after the NeurIPS Cross-Domain Meta-Learning Competition 2022.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset is considered confidential before the NeurIPS Cross-Domain Meta-Learning Competition 2022. However it will be publicly released after the challenge.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? I If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Each instance is synthesized by OmniPrint. Each instance is an image and is directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data is synthesized using the data synthesizer OmniPrint. The involved Unicode characters were manually selected from the Unicode standard, which constitutes a set of characters from several languages around the world. The involved fonts were downloaded from a manually-defined list of URLs, the downloaded fonts were then filtered by a python program in order to filter corrupted fonts. Several distortions and noises were involved, including affine and perspective transformations, random elastic transformations, natural background, foreground text filling, etc.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data is synthesized by a data synthesizer OmniPrint. The sampling is uniformly random in the given transformation parameter space.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data is synthesized by a computer software. However the design and implementation of the software, the choice of characters and fonts involve the authors of OmniPrint [11].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The datasets were synthesized on June 24, 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

N/A

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

N/A

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing/cleaning/labeling was performed. The datasets are made available as they were synthesized. No feature extraction or removal of instances was done.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing software is available in the Meta-Album Github repository and OmniPrint repository (https://github.com/SunHaozhe/OmniPrint). Details are provided on the Meta-Album website: https://meta-album.github.io/.

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Yes, a dedicated github repository will be active once the dataset is publicly released. Details are provided on the Meta-Album website (https://meta-album.github.io/). This website will also be used to announce any necessary information related to the dataset.

What (other) tasks could the dataset be used for?

Besides few-shot learning classification tasks, this datasets can be used for classification tasks of a large number of characters, and for domain adaptation tasks. Furthermore, as the meta-data can serve as labels, other kinds of classification or regression problems can also be considered e.g. classification of fonts, classification of languages, regression of rotation angle, regression of horizontal shear, etc. Finally, the datasets can be used to study disentangling the label (class character) from the nuisance variables (font, style, distortions).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

This dataset and all other datasets in Meta-Album have been prepared for competitions and benchmarks in machine learning and no other purposes. We do not make any warranties that is is appropriate for conducting scientific research other that research on machine learning algorithms nor that it is fit for developing products, whether commercial or not. In particular, this dataset may include biases that could render it unfit for such other purposes.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Until possible biases are further investigated, the dataset should not be used for any other purpose than its primary intended purpose (competitions and benchmarks).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be made available to everyone. More details about distribution can be found on the Meta-Album website (https://meta-album.github.io/).

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be released after the NeurIPS Cross-Domain Meta-Learning Competition 2022. The access information and any necessary updates will be announced via the Meta-Album website (https://meta-album.github.io/). During the review process, the dataset will be accessible to reviewers via a password-protected link.

When will the dataset be distributed?

The dataset will be distributed after the NeurIPS Cross-Domain Meta-Learning Competition 2022

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is public for research and is released with its original license : CC BY 4.0 (https://creativecommons.org/licenses/by-nc/4.0/). Further information about licenses can be found in the 'info.json' meta-data file. The license information is also mentioned on the website (https://meta-album.github.io/).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of Meta-Album paper will be responsible for supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The preferred way to contact the maintainers is to raise issues on github repository details provided on Meta Album website(https://meta-album.github.io/). In case of emergency, the authors of Meta-Album paper can be contacted via email: meta-album@chalearn.org.

Is there an erratum? If so, please provide a link or other access point.

Any necessary information or updates will be accessible via the corresponding website (https://meta-album.github.io/).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no intention to update the dataset unless required. In any case updates will be available on the website (https://meta-album.github.io/).

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Any necessary information or updates will be accessible via the website (https://meta-album.github.io/).

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

We have provided complete protocol of how such datasets can be produced in Meta-Album paper and the procedures and code for verification/validation of new constructed datasets using the defined protocol is given in the github repository (details on our website: https://meta-album.github.io/). All updates will be available on the website and the authors can be contacted via email: meta-album@chalearn.org.

Any other comments?

References

[1] G. J. Burghouts and J.-M. Geusebroek. "Material-Specific Adaptation of Color Invariant Features". In: *Pattern Recognition Letters* 30.3 (2009), pp. 306–313.

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. "Describing Textures in the Wild". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3606–3613.

[3] M. Fritz, E. Hayman, B. Caputo, and J. Eklundh. *THE KTH-TIPS database*. https://www.csc.kth.se/cvap/databases/kth-tips/index.html. 2004.

[4] G. Kylberg. *The Kylberg Texture Dataset v. 1.0*. External report (Blue series) 35. Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, 2011.

[5] S. Lazebnik, C. Schmid, and J. Ponce. "A sparse texture representation using local affine regions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1265–1278.

[6] P. Mallikarjuna, A. T. Targhi, M. Fritz, E. Hayman, B. Caputo, and J. Eklundh. *THE KTH-TIPS 2 database*. https://www.csc.kth.se/cvap/databases/kth-tips/index.html. 2006.

[7] M. Nilsback and A. Zisserman. "Automated Flower Classification over a Large Number of Classes". In: *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 2008, pp. 722–729.

[8] N. Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66.

[9] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, and N. Batra. "PlantDoc: A Dataset for Visual Plant Disease Detection". In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 2020, pp. 249–253.

[10] H. M. Sosik, E. E. Peacock, and E. F. Brownlee. *Annotated Plankton Images - Data Set for Developing and Evaluating Classification Methods*. https://hdl.handle.net/10.1575/1912/7341. 2015.

[11] H. Sun, W.-W. Tu, and I. M. Guyon. "OmniPrint: A Configurable Printed Character Synthesizer". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021.