Connecting Multi-modal Contrastive Representations

Anonymous Author(s)
Affiliation
Address
email

A Ablation Study about Text Dataset.

We conduct more experiments on audio-image retrieval with training texts from different sources. Furthermore, we provide insights about selecting training data when employing our C-MCR to connect other MCRs. As discussed in Sec 4.1, our training texts are collected from three sources: image-text datasets (COCO [1] and CC3M [2]), video-text datasets (MSRVTT [3] and MAD [4]), and audio-text datasets (AudioCap [5] and Clotho [6]).

In Table 1, we exclude the text data from image-text, video-text, and audio-text datasets, respectively. The results demonstrate that combining data from all three sources achieves the best performance. Furthermore, our findings suggest that the information in video-text datasets is relatively less important than in image-text and audio-text datasets. When applying our C-MCR to connect other MCRs, it is critical to collect overlapping modality data associated with information from non-overlapping modalities to ensure robust connections. The data from the pre-training datasets used by MCRs could serve as an appropriate starting point. Combining the overlapping modality data from these sources ensures that the data used for constructing connections contains sufficient information from non-overlapping modalities. Additionally, this data is easily accessible and scalable, which greatly enhances the practicality of our C-MCR.

Figure 1: Ablation studies about text datasets for training.

Dataset	AVE		Flickr	
	A2I	I2A	A2I	I2A
Full	**4.11**	**4.13**	**4.57**	**4.92**
w/o image-text	3.83	3.76	3.97	3.91
w/o video-text	4.04	4.05	4.41	4.78
w/o audio-text	4.07	3.88	4.31	4.59

B Downstream Task Details.

B.1 Audio-Image Retrieval.

We consider two datasets for this task: AVE [7] and Flickr-SoundNet [8], both of which consist of semantically matched image and audio pairs that were manually curated. To more comprehensively and stably reflect the retrieval capability of the model, we use all available data in these two datasets for evaluation, resulting in 4,095 samples for AVE and 5,000 samples for Flickr-SoundNet.

B.2 Audio-Visual Source Localization.

We conduct experiments on the VGGSS [9] and MUSIC [10] datasets. VGGSS is derived from VGGSound, and its test set comprises 5,158 audio-image pairs. MUSIC consists of 489 untrimmed videos of musical solos spanning 11 instrument categories for testing. It is worth noting that we use the category names from the COCO dataset as prompts to enable the open-vocabulary object detector GLIP [11] to extract object proposals.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

B.3 Counterfactual Audio-Image Recognition.

The Extended Flickr-SoundNet [12] and Extended VGGSS [12] are constructed by adding 250 and 5,158 negative samples to the test sets of the original Flickr-SoundNet and VGGSS datasets, respectively. The prompts used for the object detector GLIP [11] are also the category names from the COCO dataset. We evaluate the counterfactual Audio-Image Recognition performance using the Maximum F1 (Max-F1) and Average Precision (AP) metrics, following [12]. During inference, for the i-th image-audio pair, the proposal with the highest matching score with the audio is considered the predicted object, and its matching score is considered the confidence score c_i. The CIoU of the predicted object is denoted as IoU_i. The ground-truth map is denoted as $\mathcal{G}_i$, and the ground-truth maps of negative samples are $\emptyset$. Under these definitions, the true positives $\mathcal{TP}$, false positives $\mathcal{FP}$, and false negatives $\mathcal{FN}$ are computed as:

$$\begin{aligned}\mathcal{TP}(\gamma,\delta) &= \{i|\mathcal{G}_i \neq \emptyset, IoU_i > \gamma, c_i > \delta\}\\ \mathcal{FP}(\gamma,\delta) &= \{i|\mathcal{G}_i \neq \emptyset, IoU_i \leq \gamma, c_i > \delta\} \cup \{i|\mathcal{G}_i = \emptyset, c_i > \delta\}\\ \mathcal{FN}(\gamma,\delta) &= \{i|\mathcal{G}_i \neq \emptyset, c_i \leq \delta\}\end{aligned} \tag{1}$$

where γ is the threshold of IoU and δ is the threshold of confidence score. Following previous work, the γ is set as 0.5. The F1 score can be represented as:

$$F1(\gamma,\delta) = \frac{2 * \text{Precision}(\gamma,\delta) * \text{Recall}(\gamma,\delta)}{\text{Precision}(\gamma,\delta) + \text{Recall}(\gamma,\delta)} \tag{2}$$

where

$$\text{Precision}(\gamma,\delta) = \frac{|\mathcal{TP}(\gamma,\delta)|}{|\mathcal{TP}(\gamma,\delta)| + |\mathcal{FP}(\gamma,\delta)|};\ \ \text{Recall}(\gamma,\delta) = \frac{|\mathcal{TP}(\gamma,\delta)|}{|\mathcal{TP}(\gamma,\delta)| + |\mathcal{FN}(\gamma,\delta)|} \tag{3}$$

In accordance with [12], we calculate F1 scores for all values of δ and report the maximum F1 score (Max-F1). Average Precision (AP) is another commonly used metric in object detection, its computation is detailed in [1, 12].

C Model Configurations.

Table 1: Model configurations of projectors.

Module	Block	C_{in}	C_{out}
Projector1	Linear	512	1024
	BatchNorm1D	1024	1024
	Relu	-	-
	Linear	1024	512
	BatchNorm1D	512	512
	Relu	-	-
Projector2	Linear	512	1024
	BatchNorm1D	1024	1024
	Relu	-	-
	Linear	1024	512
	BatchNorm1D	512	512
	Relu	-	-

The model configurations of our projectors are shown in Table 1.

D Limitations and Future Work.

While C-MCR offers an efficient and effective contrastive representation learning method for modalities that lack high-quality, large-scale paired data, it still necessitates an intermediate modality to associate these modalities. Exploring ways to reduce data requirements further while maintaining representation performance is an intriguing direction for future research.

E Social Impacts.

Although C-MCR achieves outstanding performance in audio-visual learning by connecting CLIP and CLAP, further analysis of the capability boundary of this representation is necessary before applying it to additional modalities or deploying it in practice. C-MCR only requires unpaired unimodal data during training, significantly reducing the data requirements for learning a generalizable representation. However, this also means that unsuitable and harmful data in each modality are more likely to be used for training.

References

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[2] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[3] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[4] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022.

[5] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[6] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

[7] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[8] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[10] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[11] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[12] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *arXiv preprint arXiv:2209.09634*, 2022.