CyCLIP: Cyclic Contrastive Language-Image Pretraining

Shashank Goel  
UCLA  
shashankgoel@ucla.edu

Hritik Bansal  
UCLA  
hbansal@ucla.edu

Sumit Bhatia  
MDSR Lab, Adobe Systems  
sumit.bhatia@adobe.com

Ryan A. Rossi  
Adobe Research  
ryrossi@adobe.com

Vishwa Vinay  
Adobe Research  
vinay@adobe.com

Aditya Grover  
UCLA  
adityag@cs.ucla.edu

Abstract

Recent advances in contrastive representation learning over paired image-text data have led to models such as CLIP [44] that achieve state-of-the-art performance for zero-shot classification and distributional robustness. Such models typically require joint reasoning in the image and text representation spaces for downstream inference tasks. Contrary to prior beliefs, we demonstrate that the image and text representations learned via a standard contrastive objective are not interchangeable and can lead to inconsistent downstream predictions. To mitigate this issue, we formalize consistency and propose CyCLIP, a framework for contrastive representation learning that explicitly optimizes for the learned representations to be geometrically consistent in the image and text space. In particular, we show that consistent representations can be learned by explicitly symmetrizing (a) the similarity between the two mismatched image-text pairs (cross-modal consistency); and (b) the similarity between the image-image pair and the text-text pair (in-modal consistency). Empirically, we show that the improved consistency in CyCLIP translates to significant gains over CLIP, with gains ranging from $10\% - 24\%$ for zero-shot classification accuracy on standard benchmarks (CIFAR-10, CIFAR-100, ImageNet1K) and $10\% - 27\%$ for robustness to various natural distribution shifts. The code is available at https://github.com/goel-shashank/CyCLIP.

1 Introduction

The ability to learn general-purpose representations from diverse data modalities is a long-standing goal of artificial intelligence (AI) [4, 32]. In this regard, recent instantiations such as CLIP [44], ALIGN [29], and BASIC [41] have scaled up vision-language contrastive pretraining to jointly learn image and text embeddings, by exploiting an enormous amount of paired image-text data on the web. Post pretraining, these embeddings exhibit impressive zero-shot classification performance [13] and robustness to natural distribution shifts [48, 57, 24, 26]. Recently, these embeddings have been extended to text-guided generation of natural images [47, 12, 38, 46] and transferred to modalities such as 3-D shapes [50] by emphasizing the interchangeability of the image and text embeddings.

In the context of vision-language pretraining, the standard contrastive learning objective aims to maximize the similarity between matched image-text pairs (“positives”) against all the mismatched image-text pairs (“negatives”) [45, 7, 40, 22]. While such an objective aligns the true image-text pairs, it poses no constraints on the overall geometry of all data pairs, including the mismatched
10.6% on CIFAR-10 and 23.9% on CIFAR-100 respectively. Further, C with an average relative gain of when a simple nearest neighbor classifier in the text space would correctly infer the label to be a cat.

To mitigate these challenges, we propose CYCLIP. Our contributions are as follows:

1. We analyze contrastive learning for representation learning jointly over image and text modalities. We identify a critical shortcoming in the geometry of the learned representation space that can lead to inconsistent predictions in image and text domains.

Empirically, we demonstrate that the improved consistency in CYCLIP translates to improvements over CLIP. In all cases, we pre-train our models on the Conceptual Captions 3M dataset[52]. On zero-shot classification, we observe that CYCLIP improves over CLIP by 10.2% on ImageNet1K, 10.6% on CIFAR-10 and 23.9% on CIFAR-100 respectively. Further, CYCLIP outperforms CLIP with an average relative gain of +17% on ImageNet natural distribution shift benchmarks. We further analyze the improved performance of CYCLIP and find that the additional geometric structure in the representation space better captures the coarse and fine-grained concept hierarchies of datasets.

Our contributions are as follows:

Figure 1: An illustration of the planar geometry of the learned representations of image-text pairs by (a) CLIP and (b) CYCLIP. The edges indicate the distance between the representations i.e., $d(e_1, e_2) = 1 - \langle e_1, e_2 \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product. CYCLIP is cyclic consistent between image-text pairs as the in-modal distances, $d(I_{cat}, T_{dog}) \sim d(I_{cat}, I_{dog})$, and the cross-modal distances, $d(T_{cat}, I_{dog}) \sim d(I_{cat}, T_{dog})$, are similar to each other unlike CLIP. Due to explicit consistency constraints, the test image of a cat is classified as a cat in the image as well as the text space.
2. We propose CyCLIP, a simple and effective framework for contrastive representation learning with two additional cycle consistency constraints for mitigating the above issue.
3. We demonstrate that CyCLIP achieves significant empirical improvements over CLIP on zero-shot classification and robustness benchmarks. We further explain these improvements by analyzing the impact of consistency on the hierarchical structure of datasets.

2 Cycle Consistent Representation Learning

2.1 Preliminaries

We are interested in using text supervision to learn general-purpose visual representations that can be generalized to downstream predictive tasks. To this end, there have been several recent advances in language-image pretraining concerning model architectures, training objectives, and sources of supervision. Our work is most closely related to Contrastive Language-Image Pretraining (CLIP) [44] which combines many such advances in a highly scalable and generalizable learning framework.

CLIP is trained on millions of images with their captions scraped from the web. Formally, we consider a dataset \( S \subseteq \mathcal{I} \times \mathcal{T} \) consisting of pairs \((I_j, T_j)\) where \(I_j\) is a raw image and \(T_j\) is a text caption. We use \(\mathcal{I}\) and \(\mathcal{T}\) to denote the domain of images and text, respectively. The CLIP architecture consists of 3 components: (i) an image encoder network, \(f_I: \mathcal{I} \rightarrow \mathbb{R}^d\), to encode the raw image into an embedding vector of dimension \(d\); (ii) a text encoder network, \(f_T: \mathcal{T} \rightarrow \mathbb{R}^d\), to encode the raw text into an embedding vector of dimension \(d\); (iii) a contrastive objective that pulls the embeddings of paired image-caption pairs together while pushing apart embeddings of unmatched pairs.

Formally, during training, consider a batch of \(N\) image-caption pairs, \(\{I_j, T_j\}_{j=1}^N\), where \(I_j\) and \(T_j\) represent the raw image and text pair, respectively. The image embedding \(I_j^e \in \mathbb{R}^d\) and text embedding \(T_j^e \in \mathbb{R}^d\) are obtained by passing \(I_j\) and \(T_j\) through the image encoder \(f_I\) and text encoder \(f_T\), respectively; i.e. \(I_j^e = f_I(I_j)\) and \(T_j^e = f_T(T_j)\). Further, we assume they are normalized to have unit \(\ell_2\)-norm. The contrastive objective in CLIP aims to align the image and text representations by minimizing the loss function \(L_{CLIP}\) shown below:

\[
L_{CLIP} = -\frac{1}{2N} \sum_{j=1}^N \log \left( \frac{\exp \left( \langle I_j^e, T_j^e \rangle / \tau \right)}{\sum_{k=1}^N \exp \left( \langle I_k^e, T_j^e \rangle / \tau \right)} \right) - \frac{1}{2N} \sum_{k=1}^N \log \left( \frac{\exp \left( \langle I_j^e, T_k^e \rangle / \tau \right)}{\sum_{j=1}^N \exp \left( \langle I_j^e, T_k^e \rangle / \tau \right)} \right)
\]

where \(\langle \cdot, \cdot \rangle\) represents the inner product, and \(\tau\) is a trainable temperature parameter. CLIP and its variants can be used to perform zero-shot image classification, i.e., classifying test images into categories not seen at training time. We first transform each category into a suitable caption (e.g., the airplane category in CIFAR-10 can be expressed as ‘a photo of an airplane’). Then, the similarity of the test image to each caption is computed (e.g., cosine distance), and the model predicts the category for which the image-caption similarity is the highest.

2.2 Inconsistent Representation Learning in CLIP

As illustrated in Figure [a], the standard contrastive objective in CLIP can learn image-text representations such that the predicted labels for the test image are different in the image and text spaces. Here, we reason about such inconsistencies more formally in the context of downstream classification. As discussed above, we can predict a label in the text embedding space (zero-shot setting) by selecting the label that is closest to the test image \((P_T)\). Additionally, for classification in the image embedding space, if we had access to a labeled training set, then one natural way to infer the predicted label \((P_I)\) of a test image \(I_{test}\) is by taking a majority vote from the true labels associated with the k-nearest training images. Formally, we define a consistency score that measures the synchrony between the predicted labels in the image and text spaces as:
Figure 2: Illustrative overview for CyCLIP ($N = 2$). It consists of 3 major components: (a) cross-modal contrastive alignment, (b) cross-modal consistency, and (c) in-modal consistency. Only (a) is present in CLIP, whereas our proposed regularizers in (b) and (c) mitigate inconsistency.

Consistency Score $k = 1$

$$\text{Consistency Score}_k = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I} [P^k_I(I_j) = P^k_T(I_j)]$$

where $N$ is the number of test images. In our experiments (discussed in detail in §3), we found the CLIP’s consistency score ($k = 1$) to be 44%, 16%, and 16% on the standard benchmarks CIFAR-10, CIFAR-100, and ImageNet1K, respectively, showing a very high degree of disagreement in the image and text spaces. In the following section, we describe our approach to alleviate the inconsistent inference problem and quantitatively show that our solution improves the consistency score in §4.1.

2.3 Cycle Consistent Representation Learning via CyCLIP

We showed that the visual representations learned by CLIP could be inconsistent when used for inference in the image and text spaces. To mitigate this problem, we propose CyCLIP, a learning framework that builds upon CLIP by augmenting the contrastive loss in Eq. (1) with additional geometric consistency regularizers. The intuition follows directly from Figure 1 (b), where we showed that inconsistency in the image and text spaces could be eliminated if we symmetrize the similarity between the two mismatched image-text pairs and the similarity between the image-image pair and the text-text pair. We formalize this intuition with two consistency regularizers.

1. The cross-modal consistency regularizer reduces the gap in the similarity scores between the embeddings of all the mismatched image-text pairs in a batch, two at a time:

$$\mathcal{L}_{C\text{-Cyclic}} = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \left( \langle I^c_j, T^c_k \rangle - \langle I^c_k, T^c_j \rangle \right)^2.$$  

2. The in-modal consistency regularizer reduces the gap in the similarity scores between the embeddings of all combinations of image pairs and their corresponding text pairs in a batch:

$$\mathcal{L}_{I\text{-Cyclic}} = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \left( \langle I^c_j, I^c_k \rangle - \langle T^c_j, T^c_k \rangle \right)^2.$$  

Hence, our overall loss for CyCLIP is given as:

$$\mathcal{L}_{\text{CyCLIP}} = \mathcal{L}_{\text{CLIP}} + \lambda_1 \mathcal{L}_{I\text{-Cyclic}} + \lambda_2 \mathcal{L}_{C\text{-Cyclic}}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters controlling the importance of the in-modal and cross-modal cyclic consistency regularizers relative to the contrastive loss in CLIP. We can also
characterize the effect of the regularizers in terms of symmetrizing the in-modal and cross-modal similarity matrices, as illustrated in Figure 2. Note that the optimal solution to the contrastive loss formulation would push the similarity between the normalized embeddings of the matched pairs towards 1 while forcing all other pairs of similarities to 0, thereby also symmetrizing the cross-modal similarity matrix and minimizing the cross-modal consistency loss. However, this idealized scenario does not occur in practice, and we find that explicit regularization via cycle-consistency in CyCLIP facilitates improved learning, as we show in our experiments.

3 Experiments

Setup: We use Conceptual Captions 3M [52] (CC3M) image-caption pairs as the source of multimodal pretraining data for all our models. Note while this dataset is smaller than the custom dataset (400 million pairs) used in the original work on CLIP [44], it is suitable for our available data and compute and has been used for benchmark evaluations in many subsequent works on language-image pretraining [5][33][57][56]. Following prior work [44], our CLIP models use ResNet-50 as the image encoder and a transformer architecture as the text encoder. Further, we train our models from scratch for 64 epochs on 4 V100 GPUs with a batch size of 128 and an initial learning rate of 0.0005 with cosine scheduling and 10000 warmup steps. The dimension of the image and text embeddings is 1024. For CyCLIP, we use $\lambda_1 = 0.25$ and $\lambda_2 = 0.25$ across all our experiments.

3.1 Zero-Shot Transfer

We compare the zero-shot performance of CLIP and CyCLIP on standard image classification datasets: CIFAR-10, CIFAR-100 [31], and ImageNet1K [49]. We follow the evaluation strategy suggested by [44] for zero-shot classification using prompt engineering. For each dataset, we use the names of the classes to form a set of natural sentences such as ‘a photo of the {class name}’, ‘a sketch of the {class name}’ and more. These are passed through the text encoder to get a set of text embeddings for that class. This set of text embeddings are $\ell_2$-normalized, averaged, and further $\ell_2$-normalized to obtain a single text embedding for that class. For a given image, the image embedding is obtained as described in §2. The class whose text embedding (as described above) is closest to the test image is taken to be the predicted label. The zero-shot performance of the models is presented in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>ImageNet1K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top1</td>
<td>Top3</td>
<td>Top5</td>
</tr>
<tr>
<td>CLIP</td>
<td>46.54</td>
<td>78.22</td>
<td>91.16</td>
</tr>
<tr>
<td>CYCLIP</td>
<td>51.45</td>
<td>79.57</td>
<td>91.80</td>
</tr>
<tr>
<td>%GAIN</td>
<td>+10.6</td>
<td>+1.7</td>
<td>+0.7</td>
</tr>
</tbody>
</table>

We observe that the CyCLIP outperforms CLIP across all the datasets and on all TopK metrics, with gains in the range of 10% - 24% for K = 1. Our results on zero-shot transfer indicate the usefulness of having geometrical consistency for improved downstream performance of CLIP.

3.2 Robustness to Natural Distribution Shifts

One of the major successes of CLIP was its state-of-the-art performance on the natural distribution shift benchmarks. These benchmarks include images depicting sketches, cartoons, adversaries generated using attacks on trained ImageNet models. In Table 2 we evaluate the zero-shot classification accuracy of CyCLIP on four natural distribution shift benchmarks for the ImageNet dataset: ImageNetV2 [48], ImageNetSketch [57], ImageNet-A [27], and ImageNet-R [25].

For most of the distribution shift benchmarks, both CLIP and CyCLIP undergo a significant reduction in their zero-shot performance compared to the original ImageNet1K dataset (last three columns in Table 1). However, we observe that CyCLIP outperforms CLIP on all of the datasets considered in
Table 2: Zeroshot Classification on Natural Distribution Shifts (%)

<table>
<thead>
<tr>
<th></th>
<th>ImageNetV2</th>
<th>ImageNetSketch</th>
<th>ImageNet-A</th>
<th>ImageNet-R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top1</td>
<td>Top3</td>
<td>Top5</td>
<td>Top1</td>
</tr>
<tr>
<td>CLIP</td>
<td>16.91</td>
<td>29.28</td>
<td>34.99</td>
<td>10.37</td>
</tr>
<tr>
<td>CyCLIP</td>
<td><strong>19.22</strong></td>
<td><strong>32.29</strong></td>
<td><strong>38.41</strong></td>
<td><strong>12.26</strong></td>
</tr>
<tr>
<td>%GAIN</td>
<td>+13.7</td>
<td>+10.3</td>
<td>+9.8</td>
<td>+18.2</td>
</tr>
</tbody>
</table>

this experiment by a significant margin of improvement (10 - 27%). This result indicates that having cyclic consistency in the learned representations preserves the robustness on the traditional datasets.

3.3 Linear Probing

While the primary focus of CLIP and CyCLIP is zero-shot generalization, we can also assess if the benefits of our cyclic consistency constraints in mitigating inconsistency can be recovered with extra in-domain and in-modality supervision i.e., in the presence of in-distribution training samples from in-domain visual datasets. To this end, we conduct an additional experiments on linear probing where we fit a linear classifier on the representations learned by the visual encoder (ResNet-50) of CLIP and CyCLIP on a range of image classification datasets.

Table 3: Transfer CLIP and CyCLIP to 14 downstream visual datasets using linear probing. Our CyCLIP performs marginally better on 9 out of 14 datasets. For training ImageNet1K, we use a random subset of 50K images from its original training dataset.

<table>
<thead>
<tr>
<th></th>
<th>Caltech101</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>DTD</th>
<th>Aircraft</th>
<th>Flowers102</th>
<th>Food101</th>
<th>GTSRB</th>
<th>ImageNet1K</th>
<th>OxfordPets</th>
<th>SST2</th>
<th>StanfordCars</th>
<th>STL10</th>
<th>SVHN</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>79.80</td>
<td><strong>78.26</strong></td>
<td>54.85</td>
<td>59.02</td>
<td>28.00</td>
<td><strong>83.50</strong></td>
<td>54.44</td>
<td>69.72</td>
<td>35.93</td>
<td>57.66</td>
<td>53.82</td>
<td>20.00</td>
<td>89.23</td>
<td>47.28</td>
<td>57.96</td>
</tr>
<tr>
<td>CyCLIP</td>
<td><strong>80.33</strong></td>
<td>76.98</td>
<td>55.74</td>
<td>63.44</td>
<td>27.86</td>
<td>82.96</td>
<td><strong>54.96</strong></td>
<td>71.70</td>
<td><strong>37.12</strong></td>
<td>56.82</td>
<td>53.74</td>
<td>22.14</td>
<td>90.10</td>
<td>48.01</td>
<td>58.71</td>
</tr>
</tbody>
</table>

We present our results in Table 3. We find that both CLIP and CyCLIP can recover most of the performance lost due to inconsistency when provided extra in-domain and in-modality supervision, with CyCLIP marginally outperforming the CLIP on 9 out of 14 visual datasets.

4 Analysis

Previously, we demonstrated the gains of CyCLIP over CLIP on downstream tasks that involve joint reasoning over the image and text spaces. In the current section, we wish to better understand the relative behavior of the two models on a set of challenging tasks.

4.1 Consistency in Image and Text Spaces

We begin by quantitatively measuring the inconsistency problem illustrated in Figure 1. That is, we wish to evaluate to what extent are the predictions in the image-text space (zero-shot) consistent with the ones made purely within the image space, as measured by our consistency metric in Eq. 2.

Table 4 presents our results over standard benchmarks (CIFAR-10, CIFAR-100, ImageNet1K). The consistency score is calculated over 10K, 10K, and 50K testing images of the CIFAR-10, CIFAR-100 and ImageNet dataset respectively. We use 50K samples from the training set of each dataset for k-Nearest Neighbor prediction. CyCLIP is more consistent than CLIP across all the datasets as we explicitly symmetrize the cross-modal and in-modal distances. Hence, the representations learned by CyCLIP can be better used interchangeably than CLIP.
Table 4: Consistency score (%) trend for CLIP and CYCLIP across standard benchmarks. Top-k consistency score implies the fraction of times, the zero-shot predicted label in the text space is identical to the k-Nearest Neighbor predicted label in the image space (using the training dataset).

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-10</th>
<th></th>
<th>CIFAR-100</th>
<th></th>
<th>ImageNet1K</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top1</td>
<td>Top3</td>
<td>Top5</td>
<td>Top10</td>
<td>Top1</td>
<td>Top3</td>
</tr>
<tr>
<td>CLIP</td>
<td>44.60</td>
<td>46.04</td>
<td>47.06</td>
<td>48.45</td>
<td>16.21</td>
<td>17.28</td>
</tr>
<tr>
<td>CYCLIP</td>
<td>48.81</td>
<td>50.89</td>
<td>52.30</td>
<td>53.71</td>
<td>20.43</td>
<td>21.96</td>
</tr>
<tr>
<td>%GAIN</td>
<td>+8.6</td>
<td>+9.5</td>
<td>+10.0</td>
<td>+9.8</td>
<td>+20.7</td>
<td>+21.3</td>
</tr>
</tbody>
</table>

4.2 Fine-grained and Coarse-grained Performance

In §3.1, we observed that CYCLIP outperforms CLIP on zero-shot transfer across various datasets. We perform an error analysis investigating both models’ coarse and fine-grained classification performance to understand the transfer phenomena better. Given a hierarchical class structure dataset, coarse-grained classification differentiates between high-level (parent) classes, i.e., zero-shot classification into aquatic mammals and fish. The fine-grained classification task focuses on differentiating low-level (child) classes, i.e., zero-shot classification into a dolphin, otter, and seal (subclasses of aquatic mammals). We perform this analysis on the CIFAR-100, ImageNet1K, ImageNetV2, ImageNetSketch, ImageNet-A, and ImageNet-R datasets.

Formally, we consider a test set of \( N \) image-subclass-superclass triplets, \( \{I_j, C_j, P_j\}_{j=1}^N \), where \( I_j, C_j, P_j \) represent the image, the subclass (child) and superclass (parent) respectively. The image embedding \( I^e_j \in \mathbb{R}^d \) is obtained as described in §2, and the subclass embedding \( C^e_j \in \mathbb{R}^d \) and superclass embedding \( P^e_j \in \mathbb{R}^d \) are obtained as described in §3.1. Let the total number of superclasses and subclasses in the dataset be \( n_p \) and \( n_c \) respectively. Further, let \( F \) be a unique mapping from a subclass to the superclass, and \( G \) denote the inverse mapping from a superclass to the set of subclasses i.e. \( \forall P \in \{1, \ldots, n_p \}, G(P) = \{C : F(C) = P \text{ and } C \in \{1, \ldots, n_c \}\}. \) Under this setup, the fine-grained and coarse-grained accuracies are defined as:

\[
\text{Fine-grained Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I} \left( \arg\max_{C \in G(P_j)} \langle I^e_j, C \rangle = C_j \right)
\]

\[
\text{Coarse-grained Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I} \left( \arg\max_{C \in \{1, \ldots, n_c \}} \langle I^e_j, C \rangle \in G(P_j) \right)
\]

In Figure 3 we visualize how CLIP and CYCLIP compare with each other on the above metrics. The difference between the zero-shot performance of CYCLIP and CLIP is much more significant for coarse-grained classification than fine-grained classification across all the datasets. This observation indicates that concept-level knowledge is better captured in CYCLIP compared to CLIP. The drastic difference in the coarse-grained performance of CYCLIP and CLIP may be attributed to the rigid separation that the default cross-entropy loss in CLIP enforces between the positive pairs and negative pairs, which might degrade performance when some pairs in the negative batch belong to a similar entity. However, CYCLIP does not suffer from this problem as much because it poses cycle constraints on the overall geometry of all the data pairs rather than forcing a rigid separation.

4.3 Alignment and Uniformity on the Unit Hypersphere

[58] argues that contrastive learning directly optimizes for (a) alignment (closeness) of the representations of the positive pairs and (b) uniformity (coverage) of the representation space on the unit hypersphere. We extend these properties for multimodal contrastive representation learning as:

\[
\text{Alignment} = \frac{1}{N} \sum_{j=1}^{N} \langle I^e_j, T^e_j \rangle \quad \text{Uniformity} = \log \left( \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1, k \neq j}^{N} e^{-\langle I^e_j, T^e_k \rangle} \right)
\]
We desire our models to achieve high alignment and uniformity scores so that the image-text representations are close for the matched pairs and better spread over the unit hypersphere for different categories. We analyze the effect of cross-modal and in-modal consistency on the alignment and uniformity of the shared representations. For this, we train two ablated versions of CyCLIP, 1) C-CyCLIP with only cross-modal consistency component i.e. $\lambda_1 = 0, \lambda_2 = 0.5$, and 2) I-CyCLIP with only in-modal consistency component i.e. $\lambda_1 = 0.5, \lambda_2 = 0$ (in Eq. 5). We design proxy captions for classes as discussed in §3.1 to act as text embeddings. We present the results in Table 5.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>ImageNet1K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>U</td>
<td>ZS-Top1</td>
</tr>
<tr>
<td>CLIP</td>
<td>0.36</td>
<td>-0.27</td>
<td>46.54</td>
</tr>
<tr>
<td>CyCLIP</td>
<td>0.36</td>
<td>-0.34</td>
<td>51.45</td>
</tr>
<tr>
<td>I-CyCLIP</td>
<td><strong>0.60</strong></td>
<td>-0.57</td>
<td>50.97</td>
</tr>
<tr>
<td>C-CyCLIP</td>
<td>0.05</td>
<td><strong>-0.02</strong></td>
<td><strong>55.52</strong></td>
</tr>
</tbody>
</table>

We observe that I-CyCLIP learns representations that are better aligned in the representation space; however, they do not cover the hypersphere uniformly. The representations learned by C-CyCLIP are more uniformly spread but poorly aligned compared to I-CyCLIP. In this light, the components of CyCLIP can be seen to encourage a balance of good alignment and uniformity. Further, we find that CLIP is more uniform than CyCLIP in all datasets, but contrary to prior beliefs, this does not translate to improved downstream performance. CyCLIP has the best downstream zero-shot performance for CIFAR-10 and CIFAR-100 despite its poor alignment score. Further, all 3 variants of CyCLIP outperform CLIP on all 3 datasets, with CyCLIP performing the best on ImageNet1K.

### 4.4 Image-Text Retrieval

We evaluate the effectiveness of the proposed method on the cross-modal (image to text and text to image) retrieval downstream task in the zero-shot as well as fine-tuned settings. We consider the standard benchmark datasets: Flickr30K [42] and MSCOCO [8]. We assess our models on the test set of Flickr30K (1K) and MSCOCO (5K) obtained from the well-known Karpathy [30] split. Both the datasets contain 5 paired captions per image that makes text retrieval per image more easier than image retrieval per caption. We confirm the same in our results below. We perform fine-tuning on the Karpathy’s training split with the batch size of 48. We fine-tune on Flickr30K for 10 epochs and MSCOCO for 5 epochs. All the other hyperparameters are identical to that of pre-training.
Table 6 presents our cross-modal image-text retrieval results for CLIP and CyCLIP. In the zero-shot setting, we find that CyCLIP marginally outperforms CLIP on the image retrieval task on both the datasets. The relatively lower performance of both CLIP and CyCLIP in the zero-shot setting may be attributed to the more complicated nature of the two datasets where the models are expected to find similarities between the image and text at multiple resolutions as opposed to image classification where there is mostly single object to be matched with a simpler caption. It is not clear as to what distinctions in the raw input and text space are reflected in the embedding space too. Hence, we perform fine-tuning on both the datasets to better inform our models of the downstream datasets. In the fine-tuning setting, we find that the performance of both the models increases across both the datasets. However, we observe clear benefits of the soft consistent regularization on the image retrieval results for both the datasets.

4.5 CyCLIP preserves the Effective Robustness of CLIP

[36] shows that there is a strong correlation between the in-distribution and out-of-distribution generalization of the models trained on ImageNet1K, as illustrated by the linear fit (red) in Figure 4. Ideally, any model that does not undergo distribution shift would fall on the $y = x$ trendline (black). For other models, the deviations of the models from this ideal fit indicate their effective robustness. Previously, [45] showed that the zero-shot CLIP classifier trained on 400M image-text pairs improves effective robustness significantly compared to prior approaches to robustness. Subsequently, [28] demonstrated that CLIP models trained at small scales also exhibit high effective robustness that allows them to be used as a proxy to study the robustness properties of CLIP.

![Figure 4: Effect of varying the training dataset size on (a) Classification accuracy on ImageNet1K and (b) Effective Robustness on ImageNetV2.](image)

We evaluate the effect of cyclic consistency on effective robustness. We trained 4 CLIP and CyCLIP models, varying the training dataset sizes from 500K to 4M image-text pairs from the CC3M + CC12M datasets. In Figure 4 (a) we observe that for all training data sizes, CyCLIP shows a significant improvement over CLIP, showcasing its effectiveness in a diverse set of data regimes. Further, Figure 4 (b) shows that CyCLIP lies way above the baseline trend and preserves the effective robustness of CLIP.
5 Related Work

Our work fits into the broader theme of unsupervised pretraining with multiple modalities and has been successfully applied for learning representations of modalities such as images, text, and speech [2, 15, 1, 59, 43, 54]. Similar to the unimodal setting, two predominant approaches for multimodal pretraining are contrastive and generative, as described below.

Contrastive Representation Learning: Contrastive learning was originally proposed for self-supervised representation learning in the unimodal context where the embeddings of a sample are brought closer to an augmented version of the sample. In contrast, the embeddings are pushed away for other samples, and their augmentation [11, 51, 39, 55, 21, 7, 16, 40, 66, 23, 18, 63] and impose additional constraints to remove redundancies and prevent dimensional collapse in the visual representations. Recently, contrastive learning has also been used to learn robust representations of the multimodal data [62, 47]. Many works use additional losses to imbibe extra supervisory multimodal knowledge during the training process [54, 65, 64, 14, 35]. In this work, we focus on having cyclic consistency in addition to the contrastive loss to learn more robust image-text representations.

Contrastive Language-Image Pretraining: CLIP [44], ALIGN [29] and BASIC [41] have enjoyed great success in extending contrastive learning to paired image-text data, with impressive zero-shot classification and robustness performance. These works have been further extended recently to include visual self-supervision [37], additional nearest neighbor supervision [33], and utilization of unpaired data [56]. Our work complements much of this literature as it identifies consistency regularizers that can be augmented to the learning objective of the above works.

Generative Representation Learning: Generative models have been applied for learning representations of multimodal data [60, 53]. In particular, [67, 61, 10] proposed a notion of cyclic consistency for learning from unpaired multimodal data using GANs [17], which was extended later to normalizing flows [20, 19]. While these works focus on regularizing a generative mapping between modalities, our notion of cycle consistency applies to embeddings learned via a contrastive framework.

6 Conclusion

We presented CyCLIP, a framework for cycle consistent multimodal representation learning for image and text modality. The main benefits of CyCLIP stem from including cross-modal consistency and in-modal consistency regularizers to prevent inconsistent inference in the image and text spaces. Empirically, we show that CyCLIP performs much better than CLIP on zero-shot classification and is more robust on benchmarks for distributional robustness. We also showed that the representations learned by CyCLIP are more consistent than CLIP and better capture concept-level knowledge, as evidenced by our analysis of fine-grained and coarse-grained accuracies.

We believe this work can motivate further studies on understanding the geometry of the representation spaces learned via the contrastive objective applied to paired multimodal data and, in particular, identify conditions and regularization strategies under which the learned representations are synergistic across the various modalities for downstream applications.

One important future direction and a current limitation is scaling CyCLIP to larger datasets. While we do not possess the resources for this study, it is imperative to study the extent to which the benefits of cycle consistency remain at the scale on which the original CLIP was trained (400M image-text pairs). Finally, for real-world deployment of CLIP and their variants, such as CyCLIP, we need to be cautious about amplifying societal biases as these models are trained on large uncurated datasets scraped from the web [9]. Additionally, it is easy to add malicious data to the web, which poses a severe security threat [5]. Alleviating such harms is an important and active area of research.

Acknowledgements

This research is supported by an Adobe Data Science Research Award for Aditya Grover. We would like to thank the IDRE’s Research Technology group for the GPU computing resources on the UCLA Hoffman2 Cluster. We also want to thank Tung Duc Nguyen, Satvik Mashkaria, Siddarth Krishnamoorthy, Varuni Sarwal, and Ashima Suvarna for their helpful suggestions.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to extremely-compute heavy experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] All the non-proprietary datasets and code used are public under MIT, BSD or CC licenses.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]