SoftCLIP: Softer Cross-Modal Alignment Makes CLIP Stronger

Yuting Gao[1*], Jinfeng Liu[1,2*], Zihan Xu[1*], Tong Wu[1], Enwei Zhang[1], Ke Li[1], Jie Yang[2], Wei Liu[2†], Xing Sun[1†]

[1]Tencent Youtu Lab
[2]Department of Automation, Shanghai Jiao Tong University
yutinggao@tencent.com, ljf19991226@sjtu.edu.cn, ianxxu@tencent.com

Abstract

During the preceding biennium, vision-language pre-training has achieved noteworthy success on several downstream tasks. Nevertheless, acquiring high-quality image-text pairs, where the pairs are entirely exclusive of each other, remains a challenging task, and noise exists in the commonly used datasets. To address this issue, we propose SoftCLIP, a novel approach that relaxes the strict one-to-one constraint and achieves a soft cross-modal alignment by introducing a softened target, which is generated from the fine-grained intra-modal self-similarity. The intra-modal guidance is indicative to enable two pairs have some local similarities and model many-to-many relationships between the two modalities. Besides, since the positive still dominates in the softened target distribution, we disentangle the negatives in the distribution to further boost the relation alignment with the negatives in the cross-modal learning. Extensive experiments demonstrate the effectiveness of SoftCLIP. In particular, on ImageNet zero-shot classification task, using CC3M/CC12M as pre-training dataset, SoftCLIP brings a top-1 accuracy improvement of 6.8%/7.2% over the CLIP baseline.

Introduction

Since OpenAI proposed Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021), large-scale vision-language pre-training (VLP) has achieved rapid development. Many approaches (Li et al. 2021b; Yao et al. 2021; Gao et al. 2022; Li et al. 2021a) have been proposed and achieved remarkable success on several downstream tasks.

Among these methods, the alignment of the visual and linguistic modalities is a critical component, often requiring the use of image-text contrastive learning. This learning process aims to bring paired image and text samples closer while simultaneously pushing unpaired samples away, necessitating the complete mutual exclusivity between any two unpaired samples. However, acquiring high-quality image-text pairs is a challenging task, owing to the fact that the majority of image-text pairs are obtained through web crawling over the Internet, which frequently results in significant noise. As evidenced in Figure 1(a), there are some local similarities between the three pairs, the caption of (i) can also be used to describe the image (ii) and (iii), indicating many-to-many relationships instead of perfect one-to-one correspondences, which is also pointed out in CLIP-PSD (Andonian, Chen, and Hamid 2022). Therefore, it is too harsh and unreasonable to completely push away the image (i) and the text (ii)/(iii). Recent work PyramidCLIP (Gao et al. 2022) also noticed this problem and proposed to use label smoothing (Szegedy et al. 2016) to mitigate this problem. However, assigning equal weight to all the negative samples is improper and ignores the information pertaining to their relationships. The neglect of the potential distinctions among negative samples results in the underutilization of valuable information and an incomplete understanding of the underlying data structure.

In this paper, we propose SoftCLIP, a novel approach that relaxes the strict one-to-one contrastive constraint and leverages the intra-modal discriminative information to guide the interaction between visual and linguistic modalities. Specifically, we employ fine-grained intra-modal self-similarities as the softened targets for soft cross-modal alignments. Fig-

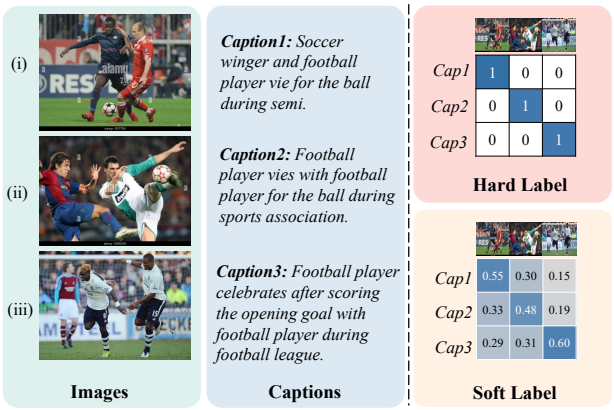

Figure 1: (a) Three image-text pairs randomly sampled from CC3M dataset have some local similarities, suggesting the ubiquitous many-to-many relationships. (b) Using fine-grained intra-modal self-similarity as the softened target can allow for the existence of some similarities among unpaired image and text.

*These authors contributed equally.
†Corresponding authors.

ure 1(b) illustrates how our softened targets allow for the existence of some similarities between the image (i) and the text (ii)/(iii). By incorporating the softened targets, SoftCLIP overcomes the limitations of traditional contrastive methods and captures the nuanced information between visual and linguistic modalities, leading to a significant improvement in cross-modal learning. Furthermore, treating different negative samples with different weights helps the model to capture the authentic distribution of data more effectively. However, the contribution of negatives in the softened target distribution can still be overwhelmed by the dominant positive one. To address this problem, we take further step to disentangle the negatives in the distribution. Specifically, we sift out the negative logits regardless of the positive logit in both prediction and target distributions with renormalization, and then bring the new two distributions closer, which boosts the relation alignment with negatives and brings further improvement.

Extensive experiments on several downstream tasks demonstrate the effectiveness of the proposed SoftCLIP. Specifically, using CC3M (Changpinyo et al. 2021)/CC12M (Sharma et al. 2018) as pre-training dataset and ResNet50 (He et al. 2016)-Transformer (Vaswani et al. 2017) as the image-text encoder, SoftCLIP achieved 24.2%/43.2% top-1 accuracy on zero-shot ImageNet (Deng et al. 2009) classification task, which is 6.8%/7.2% higher than its baseline CLIP.

Our main contributions are summarized as follows:

- We propose to employ fine-grained intra-modal self-similarities as softened targets for cross-modal learning, thereby alleviating the problem of non-strict mutual exclusion between any two pairs.
- We boost the relation alignment with negatives by disentangling the negatives in the distribution to alleviate them being overwhelmed by the positive one.
- We also use symmetric KL-Divergence to replace the conventional cross-entropy when incorporating the softened targets. Extensive experiments demonstrate the effectiveness of SoftCLIP, which can steadily bring significant improvements under various scales of pre-training data and various model architectures.

Related Work

Vision Language Pre-training

Vision-language pretraining (VLP) strives to achieve a unified representation of two modalities, namely vision and language, through the utilization of large-scale image-text pairs. Existing VLP models can be categorized into three types, *i.e.*, dual-stream models for alignment, single-stream models for fusion, or their combination.

As a paradigmatic dual-stream model, CLIP (Radford et al. 2021) has exhibited remarkable performance on zero-shot recognition and several downstream tasks by leveraging contrastive learning on large-scale image-text pairs. Following this paradigm, SLIP (Mu et al. 2022) and DeCLIP (Li et al. 2021b) further combine self-supervision to improve data utilization efficiency. PyramidCLIP (Gao et al. 2022) and FILIP (Yao et al. 2021) introduce finer-grained and more interactions between two modalities, seeking for more accurate cross-modal alignment. CyCLIP (Goel et al. 2022) points out the importance of geometric consistency in the learned representation space between two modalities, and proposes geometrically consistency constraints. Different from dual-stream ones, single-stream models, such as Visual-BERT (Li et al. 2019) and OSCAR (Li et al. 2020), fuse the image and text features with a unified model to achieve deeper interaction. ALBEF (Li et al. 2021a) and CoCa (Yu et al. 2022) absorb the essence of the two kinds of structures, and find a more flexible way to learn visual and linguistic representations. In this paper, we adopt the dual-stream architecture and depart from the commonly used one-hot labels. Instead, we utilize fine-grained intra-modal self-similarities as softened targets to provide more informative guidance, which leads to improved cross-modal interactions.

Softened Target

Softened target aims to alleviate the strict constraints imposed by one-hot label and avoid the model's overconfidence towards wrong predictions, which has demonstrated its effectiveness across various tasks. For example, label smoothing (Szegedy et al. 2016), a commonly used strategy in classification task, assigns some small positive values to the ground-truth of all negative samples. Moreover, in the field of knowledge distillation (Hinton et al. 2015), the logits predicted by the teacher model will be used as softened targets to guide the learning process of student model. The softened targets, containing the teacher's modeling of the relationship among all the samples, are more instructive than the one-hot label. Recently, PyramidCLIP (Gao et al. 2022) has pointed out the potential limitation of the overly rigid one-hot label, and hence proposes to use label smoothing to mitigate this problem. However, it should be emphasized that the indiscriminate treatment towards all negative samples is unreasonable and necessitates further attention. CLIP-PSD (Andonian, Chen, and Hamid 2022) also utilizes softened targets obtained from a teacher model to reduce the adverse effects of noisy image-text pairs. Its core concept is progressive self-distillation where the student network acts as its own teacher and the model dynamically evolves into its own teacher as training progresses. From this perspective, SoftCLIP is also working under the self-distillation framework, however, the softened targets do not stem from the images and texts, but from the pre-extracted ROI (region-of-interest) features of objects and corresponding tags.

Methodology

In this section, we first present some CLIP preliminaries, and then introduce the details of our proposed SoftCLIP. The overall framework can be seen in Figure 2.

CLIP Preliminaries and Label Smoothing

Consider a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, CLIP employs a dual-stream encoder to obtain the semantic representation of each pair. Specifically, for the i_{th} pair, the

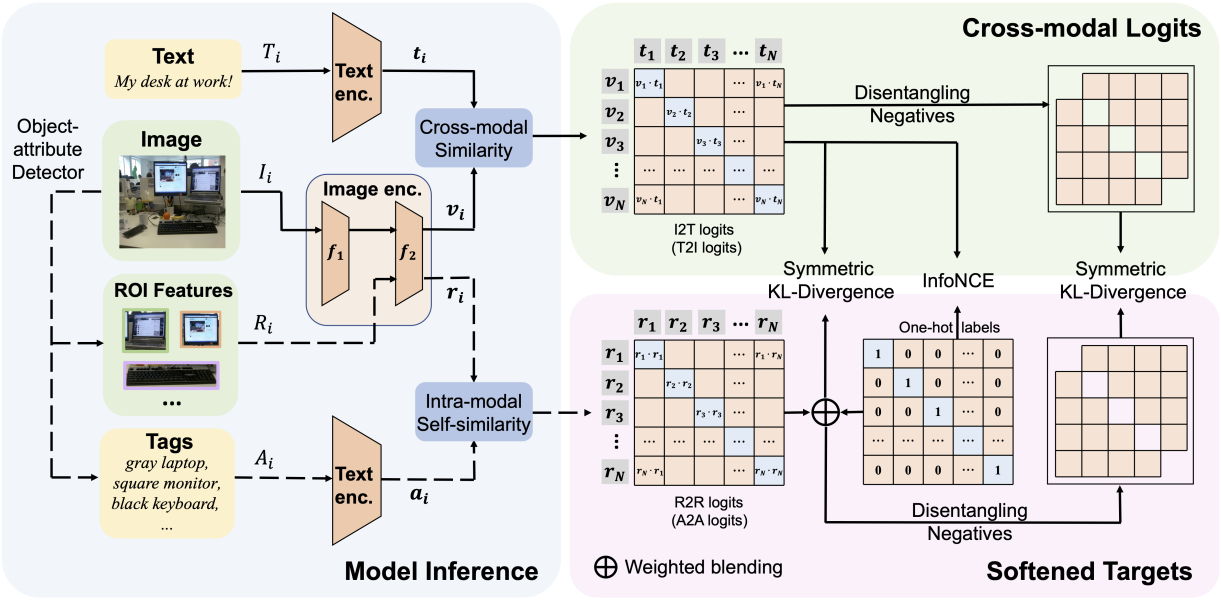

Figure 2: The overall framework of SoftCLIP. For each image-text pair, the image is fed into a pre-trained object-attribute detector to extract ROI features and their corresponding tags, which are used to compute the intra-modal self-similarities to guide the cross-modal interactions. Besides, we disentangle negatives in each distribution to construct another soft loss term and boost the relation alignment with negatives. And the conventional cross-entropy is replaced by symmetric KL-Divergence when incorporating the softened targets.

image data I_i is input into an image encoder to get the visual representation $\boldsymbol{v}_i$, and the text data T_i is input into a text encoder to get the linguistic representation $\boldsymbol{t}_i$, generating L2-normalized embedding pairs $\{(\boldsymbol{v}_i, \boldsymbol{t}_i)\}_{i=1}^N$. CLIP uses InfoNCE (Oord, Li, and Vinyals 2018) to conduct cross-modal alignment, which pulled the paired image and text embeddings together while pushing unpaired apart. For the i_{th} pair, the normalized image-to-text similarity vector $\boldsymbol{p}_i(I,T) = \{p_{ij}(I,T)\}_{j=1}^N$ and the text-to-image counterpart $\boldsymbol{p}_i(T,I) = \{p_{ij}(T,I)\}_{j=1}^N$ can be calculated through:

$$p_{ij}(I,T) = \frac{\exp(\text{sim}(\boldsymbol{v_i}, \boldsymbol{t_j})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\boldsymbol{v_i}, \boldsymbol{t_j})/\tau)}, \tag{1}$$

$$p_{ij}(T,I) = \frac{\exp(\text{sim}(\boldsymbol{t_i}, \boldsymbol{v_j})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\boldsymbol{t_i}, \boldsymbol{v_j})/\tau)}, \tag{2}$$

where τ is a learnable temperature parameter initialized with 0.07 and the function $\text{sim}(\cdot)$ conducts dot product to measure the similarity scores. In CLIP paradigm, the corresponding one-hot label vectors are used as the targets to calculate InfoNCE loss. The one-hot label of the i_{th} pair is denoted as $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^N$, with y_{ii} equal to 1 and all other elements equal to 0. Therefore the vision-to-language loss and the language-to-vision loss can be obtained by:

$$\mathcal{L}_{v2l} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{y}_i, \boldsymbol{p}_i(I,T)), \tag{3}$$

$$\mathcal{L}_{l2v} = \frac{1}{N}\sum_{i=1}^{N} H(\boldsymbol{y}_i, \boldsymbol{p}_i(T,I)), \tag{4}$$

where $H(\cdot,\cdot)$ denotes the cross-entropy operation. And the final CLIP loss can be denoted as $\mathcal{L}_{\text{CLIP}} = (\mathcal{L}_{v2l} + \mathcal{L}_{l2v})/2$.

As we have discussed, CLIP neglects some local similarities between unpaired images and texts within a batch, while PyramidCLIP roughly uses label smoothing to soften the hard one-hot targets to alleviate this issue. Specifically, the original one-hot label vector $\boldsymbol{y}_i$ is softened to $\widetilde{\boldsymbol{y}}_i$, which is formulated as:

$$\widetilde{\boldsymbol{y}}_i = (1-\alpha)\boldsymbol{y}_i + \frac{\alpha}{N-1}(\boldsymbol{1} - \boldsymbol{y}_i), \tag{5}$$

where α is the smoothing hyper-parameter set to 0.2, and $\boldsymbol{1}$ denotes the all-ones vector.

Soft Alignment under Intra-modal Guidance

The label smoothing strategy transfers a small portion of the confidence from the positive sample and amortizes it to the negatives, allowing for weak and fixed similarity with negatives. This strategy works in PyramidCLIP, however, the improvement it brings is limited since it merely models naive many-to-many relationships between images and the corresponding texts. To improve this, we try to find clues from the relation within a single modality. Specifically, we attempt to use the intra-modal self-similarity as the softened target to guide the CLIP model. An accurate intra-modal self-similarity can provide a superb supervision to repair a sample with more semantically similar correspondences in another modality. Moreover, it inherently contains the implicit expression of many-to-many relationships, with rich and instructive knowledge.

Intuitively, we may choose the original images and texts to calculate the intra-modal self-similarity, *i.e.*, image-to-image similarity for the visual modality and text-to-text similarity for the textual modality. However, this approach encounters some problems and does not perform well in practice, which is revealed in the experimental part. PyramidCLIP pre-extracts the ROI features of detected salient objects for each image, with tag description for each object, to introduce cross-level relation alignment, which can bring significant gains. The ROI features and corresponding tags of objects, extracted by a pre-trained object-attribute detector, contain the prior category and attribute information of objects from the task of object detection. This encourages us to exploit the priors, *i.e.*, we can alternatively use the ROI features and tags to calculate the intra-modal self-similarity.

Formally, for the image-text pair (I_i, T_i), we can pre-extract the corresponding ROI-tag (ROI features and tags) pair (R_i, A_i) from the image I_i, constructing ROI-tag pairs $\{(R_i, A_i)\}_{i=1}^{N}$ within a batch. Note that the tags are concatenated and separated by commas to form a sentence. Each pair is feed into the dual-stream model following PyramidCLIP. As shown in Figure 2, R_i is processed by the rear part of the image encoder and A_i is processed by the text encoder, deriving the corresponding L2-normalized representation vector pairs $\{(\boldsymbol{r}_i, \boldsymbol{a}_i)\}_{i=1}^{N}$. And the linear embedding layers for transforming vector dimension are omitted here. For the i_{th} pair, the normalized intra-modal self-similarity vectors of R_i and A_i, denoted as $\boldsymbol{p}_i(R,R) = \{p_{ij}(R,R)\}_{j=1}^{N}$ and $\boldsymbol{p}_i(A,A) = \{p_{ij}(A,A)\}_{j=1}^{N}$ respectively, can be obtained by:

$$p_{ij}(R,R) = \frac{\exp(\text{sim}(\boldsymbol{r_i}, \boldsymbol{r_j})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\boldsymbol{r_i}, \boldsymbol{r_j})/\tau)}, \tag{6}$$

$$p_{ij}(A,A) = \frac{\exp(\text{sim}(\boldsymbol{a_i}, \boldsymbol{a_j})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\boldsymbol{a_i}, \boldsymbol{a_j})/\tau)}. \tag{7}$$

Next, the ROI self-similarity and tag self-similarity are utilized as the soft labels to supervise the image-to-text and text-to-image correspondences respectively. In practice, we use the weighted average of the hard labels and the soft labels as the final softened targets to ensure the training stability and better generalization, which is formulated as:

$$\widetilde{\boldsymbol{p}}_i(R,R) = (1-\beta)\boldsymbol{y}_i + \beta \boldsymbol{p}_i(R,R), \tag{8}$$

$$\widetilde{\boldsymbol{p}}_i(A,A) = (1-\beta)\boldsymbol{y}_i + \beta \boldsymbol{p}_i(A,A), \tag{9}$$

where $\boldsymbol{y}_i$ denotes the hard one-hot label and β is a mixing coefficient set to 0.3. Since the softened targets are also variable distributions, the cross-entropy in CLIP should be replaced by the KL-Divergence as follows:

$$\mathcal{L}_{\text{soft-}v2l} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}(\widetilde{\boldsymbol{p}}_i(R,R) \,||\, \boldsymbol{p}_i(I,T)), \tag{10}$$

$$\mathcal{L}_{\text{soft-}l2v} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}(\widetilde{\boldsymbol{p}}_i(A,A) \,||\, \boldsymbol{p}_i(T,I)). \tag{11}$$

Then we can get the average soft loss under the guidance of ROIs and tags, denoted as $\mathcal{L}_{\text{soft}} = (\mathcal{L}_{\text{soft-}v2l} + \mathcal{L}_{\text{soft-}l2v})/2$.

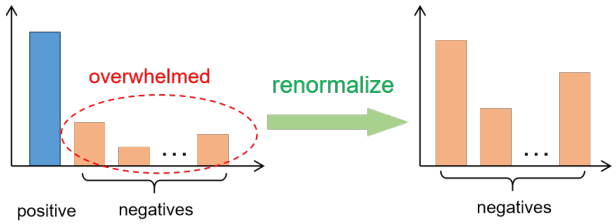

Figure 3: Disentangling the negatives in the distribution.

Boosting Relation Alignment with Negatives

The introducing of intra-modal self-similarity does relax the strict one-to-one constraint and guide the model to learn many-to-many correspondences between the visual and linguistic modalities. However, the confidence of the positive sample still dominates compared to the negatives despite of the softened target distribution. This may lead to numerous negatives submerged by the dominant positive ones in the cross-modal relation alignment. And the problem will be more serious when meeting faulty positives, which means the paired images and texts in the web-harvested dataset are actually irrelevant. To mitigate this issue, we disentangle negatives in the distribution to boost the relation alignment with negatives in SoftCLIP.

Specifically, we discard the positive logits in the probability distribution and only concentrate on the knowledge among negative logits with renormalization, as shown in Figure 3. For any distribution vector $\boldsymbol{p}_i = \{p_{ij}\}_{j=1}^{N} \in \mathbb{R}^{1\times N}$, we use $\boldsymbol{p}_i^* = [p_{i1}^*, ..., p_{i(i-1)}^*, p_{i(i+1)}^*, ..., p_{iN}^*] \in \mathbb{R}^{1\times(N-1)}$ to denote its corresponding neg-disentangled distribution, with the elements calculated through:

$$p_{ij}^* = \frac{p_{ij}}{\sum_{k=1,k\neq i}^{N} p_{ik}}, \tag{12}$$

where j is taken from $[1, ..., i-1, i+1, ..., N]$. The disentangling of negatives is applied identically to the distributions $\widetilde{\boldsymbol{p}}_i(R,R)$, $\widetilde{\boldsymbol{p}}_i(A,A)$, $\boldsymbol{p}_i(I,T)$ and $\boldsymbol{p}_i(T,I)$, generating $\widetilde{\boldsymbol{p}}_i^*(R,R)$, $\widetilde{\boldsymbol{p}}_i^*(A,A)$, $\boldsymbol{p}_i^*(I,T)$ and $\boldsymbol{p}_i^*(T,I)$ correspondingly. Then we can derive the relation-enhanced formulation of $\mathcal{L}_{\text{soft-}v2l}$ and $\mathcal{L}_{\text{soft-}l2v}$ as:

$$\mathcal{L}_{\text{soft-}v2l}^{\text{re}} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}(\widetilde{\boldsymbol{p}}_i^*(R,R) \,||\, \boldsymbol{p}_i^*(I,T)), \tag{13}$$

$$\mathcal{L}_{\text{soft-}l2v}^{\text{re}} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}(\widetilde{\boldsymbol{p}}_i^*(A,A) \,||\, \boldsymbol{p}_i^*(T,I)). \tag{14}$$

Hence, the relation-enhanced soft loss can be written as $\mathcal{L}_{\text{soft}}^{\text{re}} = (\mathcal{L}_{\text{soft-}v2l}^{\text{re}} + \mathcal{L}_{\text{soft-}l2v}^{\text{re}})/2$.

Training Objective

It is well known that the KL-Divergence is essentially asymmetric, whereas the JS-Divergence is an alternative with symmetric form. However, we have observed that the JS-Divergence makes the training stage unstable. Therefore,

we directly symmetrize the KL-Divergence by adding a reversed term with the two input distributions exchanged, which has been proved to be effective in the experiments. For instance, the symmetric form of $D = \mathrm{KL}(\boldsymbol{p} \,||\, \boldsymbol{q})$ can be written as:

$$\widetilde{D} = \frac{1}{2}(\mathrm{KL}(\boldsymbol{p} \,||\, \boldsymbol{q}) + \mathrm{KL}(\boldsymbol{q} \,||\, \boldsymbol{p})). \tag{15}$$

Following this, we can symmetrize $\mathcal{L}_{\mathrm{soft}}$ and $\mathcal{L}^{\mathrm{re}}_{\mathrm{soft}}$, obtaining $\widetilde{\mathcal{L}}_{\mathrm{soft}}$ and $\widetilde{\mathcal{L}}^{\mathrm{re}}_{\mathrm{soft}}$ respectively. And we utilize the two terms to regulate the original CLIP loss. So the overall loss function is denoted as:

$$\mathcal{L}_{\mathrm{SoftCLIP}} = \widetilde{\mathcal{L}}_{\mathrm{soft}} + \lambda \widetilde{\mathcal{L}}^{\mathrm{re}}_{\mathrm{soft}} + \mu \mathcal{L}_{\mathrm{CLIP}}, \tag{16}$$

where the loss weights λ and μ are set to 1.0 and 0.5 in the experiments.

Experiments

Pre-training and Evaluation Details

Architectures and Pre-training Datasets SoftCLIP accommodates three distinct model architectures, with the visual encoder compatible with ResNet50, ViT-B/32 (Dosovitskiy et al. 2020) and ViT-B/16 (Dosovitskiy et al. 2020), while the language encoder utilizes Transformer following CLIP (Radford et al. 2021). The input resolution of image encoder is 224×224 and the maximum context length of text encoder is 77. And SoftCLIP is pre-trained on three datasets, CC3M (Changpinyo et al. 2021), CC12M (Sharma et al. 2018) and YFCC15M-V2 (Li et al. 2021b). These datasets are listed in Table 1.

Object-attribute Detector The object-attribute detector used to extract ROI features with tags is pre-trained by VinVL (Zhang et al. 2021), adopting the framework of Faster R-CNN (Ren et al. 2015). Through the detector, we take 10 objects with the highest confidence from each image to obtain the corresponding ROI features and category descriptions with attribute information. Each ROI feature is of 2052-dimension, concatenated by a 2048-dimensional appearance feature vector and 4-dimensional position vector (the coordinates of top-left and bottom-right corners of the object region).

Implementation Details We train our SoftCLIP using an AdamW (Loshchilov and Hutter 2017) optimizer and the cosine learning rate scheduler with a linear warm-up. Specifically, the learning rate linearly increases from 0 to the peak value within 10% of the total steps, and then decreases with a cosine anneal strategy. The weight decay rate of AdamW is set to 0.2. To save GPU memory, automatic mixed-precision (Micikevicius et al. 2018) is used. The models are trained from scratch for either 8 or 32 epochs in our experiments, *i.e.*, 8 epochs for ablation and 32 epochs for comparison. We use 8 V100 GPUs for experiments, when training with ResNet50 and ViT-B/32 image encoder, the batch size is set to 2048, while with the image encoder ViT-B/16, the batch size is 1024.

Dataset	CC3M	CC12M	YFCC15M-V2
Size	3M	10M	15M

Table 1: Pre-training datasets.

Method	**Pretrain Dataset**	**Image Encoder**	**ImageNet ZS Top1**
CLIP♢	CC3M	ResNet50	17.7
SoftCLIP	CC3M	ResNet50	**24.2**
CLIP♢	CC3M	ViT-B/32	11.9
SoftCLIP	CC3M	ViT-B/32	**13.3**
CLIP♢	CC3M	ViT-B/16	16.9
SoftCLIP	CC3M	ViT-B/16	**18.9**
CLIP♢	CC12M	ResNet50	36.0
SoftCLIP	CC12M	ResNet50	**43.2**
CLIP♢	CC12M	ViT-B/32	31.5
SoftCLIP	CC12M	ViT-B/32	**34.4**
CLIP♢	CC12M	ViT-B/16	36.8
SoftCLIP	CC12M	ViT-B/16	**42.1**
CLIP♢	YFCC15M-V2	ResNet50	39.6
SoftCLIP	YFCC15M-V2	ResNet50	**43.7**
CLIP♢	YFCC15M-V2	ViT-B/32	33.1
SoftCLIP	YFCC15M-V2	ViT-B/32	**35.0**
CLIP♢	YFCC15M-V2	ViT-B/16	38.9
SoftCLIP	YFCC15M-V2	ViT-B/16	**42.4**

♢ Our Implementation

Table 2: Comparison against CLIP baseline on ImageNet Zero-Shot (ZS) classification.

Method	**Image Encoder**	**PETS**	**DTD**	**F101**	**FLOW**	**SUN**	**CAL**	**AVG**
CLIP♢	ResNet50	33.3	22.8	48.0	54.9	50.0	65.6	45.8
SoftCLIP	ResNet50	**34.9**	**27.1**	**50.8**	**56.3**	**55.9**	**70.4**	**49.2**
CLIP♢	ViT-B/16	27.2	21.6	48.3	53.8	53.4	71.5	46.0
SoftCLIP	ViT-B/16	**32.5**	**25.6**	**53.8**	**55.6**	**56.2**	**71.8**	**49.2**

♢ Our Implementation

Table 3: Accuracy on 6 datasets with ResNet50 and ViT-B/16 image encoder pretrained on YFCC15M-V2. PETS / DTD / F101 / FLOW / SUN / CAL are abbreviations for Pets / Describable Textures / Food-101 / Flowers-102 / SUN397 / Caltech-101 datasets. AVG represents average accuracy across all 6 datasets.

Downstream Tasks for Evaluation We validate the effectiveness of the proposed SoftCLIP on three downstream tasks: zero-shot image classification, zero-shot image-text retrieval and image retrieval. For zero-shot image classification, experiments are carried out on 7 datasets, such as ImageNet (Deng et al. 2009), Pets (Parkhi et al. 2012), Describable Textures (Cimpoi et al. 2014), Food-101 (Bossard, Guillaumin, and Van Gool 2014), Flowers-102 (Nilsback and Zisserman 2008), SUN397 (Xiao et al. 2010) and Caltech-101 (Fei-Fei, Fergus, and Perona 2004). For zero-shot image-text retrieval, experiments are conducted on Flickr30K (Hodosh, Young, and Hockenmaier 2013) and

Method	Image Encoder	Flickr30K(1K)						MS-COCO(5K)					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP♢	ResNet50	54.9	81.6	90.5	37.1	65.0	75.0	29.4	54.8	66.1	18.9	40.7	52.5
DECLIP†		58.7	85.0	92.5	40.7	68.9	78.4	31.1	59.0	69.9	20.6	43.8	55.4
SoftCLIP		**62.1**	**86.4**	**93.0**	**43.0**	**71.0**	**80.3**	**36.0**	**61.2**	**72.3**	**22.2**	**45.8**	**57.3**
CLIP♢	ViT-B/16	54.9	80.0	88.4	**37.2**	**64.3**	74.3	30.7	**56.2**	67.4	19.1	40.9	52.5
SoftCLIP		**56.2**	**82.1**	**88.6**	**37.2**	**64.3**	**74.5**	**30.9**	**56.2**	**68.3**	**19.2**	**41.2**	**52.6**

♢ Our Implementation

† Tested with: https://github.com/Sense-GVT/DeCLIP#supported-models

Table 4: Zero-shot image-text retrieval results on Flicker30K and MS-COCO. All models are pre-trained on YFCC15M-V2.

MS-COCO (Lin et al. 2014). For image retrieval, two subtasks are included: instance retrieval task on Oxford (Philbin et al. 2007) and Paris Buildings datasets (Philbin et al. 2008), and copy detection task on the INRIA Copydays (Douze et al. 2009) dataset. The results of image retrieval can be seen in the supplementary materials.

Zero-shot Image Classification

To validate the effectiveness of the proposed SoftCLIP, we first conduct experiments on the widely used zero-shot ImageNet classification task. The results are presented in Table 2. It is clear that SoftCLIP brings significant improvement compared to the CLIP baseline with different image encoders, across varying levels of pre-training data. Notably, SoftCLIP exhibits a significant increase of 6.5%/7.2% in top-1 accuracy compared to CLIP when the pre-training dataset is CC3M/CC12M and the visual encoder is ResNet50. Besides, we also provide the zero-shot classification results on the other six small datasets, which are illustrated in Table 3. Obviously, the performance of SoftCLIP significantly exceed the CLIP baseline across all the six datasets, which demonstrates the efficacy and generalization of the proposed SoftCLIP.

Zero-shot Image-text Retrieval

Next, we validate the efficacy of our proposed method on image-text retrieval task. To this end, we conduct zero-shot image-text retrieval experiments on the Flikcer30K and MS-COCO datasets, and present the obtained results in Table 4. The experimental results demonstrate that SoftCLIP confers significant improvements on both datasets. In particular, when the image encoder is ResNet50, SoftCLIP brings a top-1 hit accuracy improvement of 7.2% and 5.9% on Flicker30K image-to-text and text-to-image retrieval tasks respectively. Furthermore, SoftCLIP outperforms DeCLIP pre-trained with the same dataset by a significant margin.

Ablation Study

In this section, we first conduct ablation studies to demonstrate the effectiveness of each module in SoftCLIP, and then explore some other factors which may influence the performance. All the ablation experiments are conducted on CC3M for 8 epochs. More ablation results can be seen in the supplementary materials.

Method	ResNet50 IN ZS Top-1	ViT-B/32 IN ZS Top-1
CLIP (Baseline)	16.5	10.7
+ Label Smoothing	18.3	11.2
CLIP + Soft Loss	20.5	11.7
+ Relation-enhanced Soft Loss	21.4	12.2
+ Symmetric KL (SoftCLIP)	**22.1**	**12.5**

Table 5: The effectiveness of each component in SoftCLIP.

Effectiveness of Each Module To verify the effectiveness of each component proposed in SoftCLIP, we conduct a series of experiments with all components added to the CLIP paradigm successively. As demonstrated in Table 5, only the CLIP loss plus the naive soft loss $\mathcal{L}_{\text{soft}}$ can bring significant gains, even exceeding the label smoothing strategy appreciably. Moreover, the adjunction of relation-enhanced soft loss $\mathcal{L}_{\text{soft}}^{\text{re}}$ and the symmetrization of KL-Divergence can further improve the model performance.

Ablation about the Source of Softened Targets As we have mentioned in the methodology part, image and text self-similarities are more intuitive to serve as the softened targets compared with ROI and tag self-similarities. Here we provide experimental basis to demonstrate why we choose ROIs and tags. Let $\mathcal{L}(R, A)$ denote the soft loss plus relation-enhanced soft loss under the guidance of ROI and tag self-similarities, and $\mathcal{L}(I, T)$ denote that under the guidance of image and text self-similarities. We additionally experiment with a mixed loss function denoted as $\mathcal{L} = \gamma\mathcal{L}(R, A) + (1 - \gamma)\mathcal{L}(I, T)$, where γ is adjustable to control the proportion of the two terms and the CLIP loss is not included in this ablation. The variety of the model performance with respect to γ is depicted in Figure 5(a), which reveals that the model performs better with higher ratio of $\mathcal{L}(R, A)$, *i.e.*, the guidance from ROI and tag self-similarities. We attribute it to two reasons: One is that the image and text similarities are inaccurate in the early training stage, while ROIs and tags inherently contain fine-grained internal alignment due to the priors from the task of object detection; The second reason is that complete images

Figure 4: (a) Text-to-image retrieval examples on MS-COCO dataset. From left to right are the top 10 retrieved images from rank1 to rank10. (b) Grad-CAM heatmaps for finding the correspondence from word in the caption to region in the image.

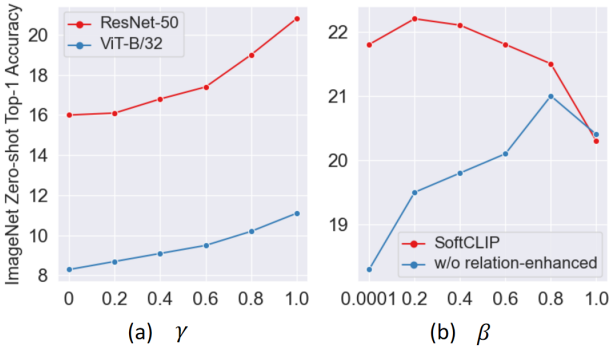

Figure 5: (a) The influence of ROI-tag guidance and image-text guidance at different mixing ratios. (b) The influence of soft self-similarity label and hard one-hot label at different mixing ratios.

and captions only provide a global understanding, which is relatively coarse, whereas ROIs and tags can capture more detailed local information, providing better guidance.

Influence of the Parameter β Recall that β is the weighting coefficient to mix the one-hot hard label and the soft self-similarity label in Equation (8) (9). Higher value of β indicates higher proportion of the self-similarity label. Here we explore the influence of β, which is shown in Figure 5(b). In SoftCLIP (see the blue line), the optimal performance is achieved with β between 0.1 and 0.5. However, as it increases to $\beta > 0.8$, the performance declines dramatically, which implies that pure self-similarity labels have very poor guidance, hence requiring the reconciliation of hard labels. Another interesting phenomenon is that the accuracy only drops slightly when we mix a very small ratio of the soft label, *i.e.*, $\beta = 0.0001$. Our explanation is that the relation-enhanced soft loss term is taking effect. A very small value of β (0.0001) leads to a dominant positive logit (more than 0.9999) in the softened target with all the negatives overwhelmed. Nevertheless, the negative logits can be prominent again after being disengaged in the distribution, hence, the model can still capture the relation with negatives. To verify this, we conduct additional experiments with the relation-enhanced soft loss removed (see the orange line). In this configuration, the model performance drops sharply when $\beta < 0.2$, which is consistent with the theoretical analysis.

Visualization

Text-to-Image Retrieval In Figure 4(a), we give some text-to-image top 10 retrieval results on MS-COCO. It can be seen in the first example that, CLIP tends to narrowly focus on the unitary and specific expression, such as "`black and white`", while ignoring others like "`birds`", resulting in the retrieval of mostly images of zebras. Whereas, SoftCLIP has a more comprehensive understanding of the text-image relationship and can retrieve the images that have a better match with the query text.

Word-level Localization Grad-CAM (Selvaraju et al. 2017) is utilized to show the word-level localization in the image for an image-text pair. As shown in Figure 4(b), SoftCLIP has more precise responses to some nouns compared to CLIP and can accurately locate the region related to the noun. For instance, in the second example, SoftCLIP can exactly locate the corresponding regions of "`cars`" and "`boat`", while CLIP are confused. We attribute this to the introduction of fine-grained softened target, *i.e.*, the object-level intra-modal self-similarity.

Conclusions

In this paper, we propose SoftCLIP, a novel approach that relaxes the strict one-to-one constraint and achieves a soft cross-modal alignment by introducing intra-modal self-similarity as softened target and disentangling negatives in the distribution. SoftCLIP can model the commonly existing many-to-many relationships in the web-crawled noisy image-text datasets. Extensive experiments on several tasks demonstrate the effectiveness of the proposed SoftCLIP.

References

Andonian, A.; Chen, S.; and Hamid, R. 2022. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16430–16441.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. European Conf. Computer Vision*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; ; and Vedaldi, A. 2014. Describing Textures in the Wild. In *Proceedings of the IEEE international conference on computer vision*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Douze, M.; Jégou, H.; Sandhawalia, H.; Amsaleg, L.; and Schmid, C. 2009. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 1–8.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE international conference on computer vision workshop*, 178–178. IEEE.

Gao, Y.; Liu, J.; Xu, Z.; Zhang, J.; Li, K.; and Shen, C. 2022. PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. *arXiv preprint arXiv:2204.14095*.

Goel, S.; Bansal, H.; Bhatia, S.; Rossi, R. A.; Vinay, V.; and Grover, A. 2022. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. European Conf. Computer Vision*, 121–137.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2021b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. European Conf. Computer Vision*, 740–755.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; and Wu, H. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, 529–544. Springer.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *Proceedings of the IEEE international conference on computer vision*, 3498–3505. IEEE.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Proc. Advances in Neural Information Processing Systems*, 28.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. 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)*, 2556–2565.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proc. IEEE/CVF Conf. Computer Vision & Pattern Recognition*, 5579–5588.