iCLIP: Bridging Image Classification and Contrastive Language-Image Pre-training for Visual Recognition

Yixuan Wei\textsuperscript{1,2}, Yue Cao\textsuperscript{2}, Zheng Zhang\textsuperscript{2}, Houwen Peng\textsuperscript{2}, Zhuliang Yao\textsuperscript{1,2}, Zhenda Xie\textsuperscript{1,2}
Han Hu\textsuperscript{2}, Baining Guo\textsuperscript{2}
\textsuperscript{1}Tsinghua University \textsuperscript{2}Microsoft Research Asia

Abstract

This paper presents a method that effectively combines two prevalent visual recognition methods, i.e., image classification and contrastive language-image pre-training, dubbed iCLIP. Instead of naive multi-task learning that use two separate heads for each task, we fuse the two tasks in a deep fashion that adapts the image classification to share the same formula and the same model weights with the language-image pre-training. To further bridge these two tasks, we propose to enhance the category names in image classification tasks using external knowledge, such as their descriptions in dictionaries. Extensive experiments show that the proposed method combines the advantages of two tasks well: the strong discrimination ability in image classification tasks due to the clean category labels, and the good zero-shot ability in CLIP tasks ascribed to the richer semantics in the text descriptions. In particular, it reaches 82.9% top-1 accuracy on IN-1K, and meanwhile surpasses CLIP by 1.8%, with similar model size, on zero-shot recognition of Kornblith 12-dataset benchmark. The code and models are publicly available at \url{https://github.com/weiyx16/iCLIP}.

1. Introduction

Image classification is a classic visual problem whose goal is to classify images into a fixed set of pre-defined categories. For example, the widely used ImageNet dataset \cite{russakovsky2015imagenet} carefully annotated 14 million images and categorize them into 21,841 categories chosen from the WordNet \cite{miller1995wordnet}. For image classification, each category provides a clear taxonomy that groups images of the same category together and separates images from different categories, and thus endows the learnt representation with strong discriminant ability. However, this classification ability is limited to a fixed set of categories \cite{russakovsky2015imagenet, auron2018voc, xiong2020deformable}.

\footnote{Corresponding Author. The work is done when Yixuan Wei, Zhuliang Yao, and Zhenda Xie are interns at Microsoft Research Asia.}

Recently, the method that learns to contrast image-text pairs, known as contrastive language-image pre-training (abbr. CLIP), has well made up such shortage of the conventional image classification methods to achieve strong zero-shot recognition ability \cite{tian2021contrastive, ramesh2021zero}. These methods employ a contrastive learning framework, where images and their corresponding alt-texts are treated as positive pairs, while images with all other alt-texts are treated as negative pairs. Thanks to the rich semantics involved in the alt-texts, the images can be weakly connected to almost arbitrary categories that already appear in the alt-texts, resulting in its zero-shot ability. A drawback is that the image-text pairs are usually crawled from the internet without human labeling, leading to their noisy and ambiguous nature. Thus the learnt representations are often not conceptual compact, and may lack certain discriminative ability.
This paper explores how to effectively combine these two powerful visual recognition and representation learning methods, to take advantages of both methods and data sources while relieving their shortages. We first try a naive multi-task learning framework that applies the original head networks of the two tasks on top of a shared visual encoder, and jointly learn the network with separate losses of the two tasks on top of a shared visual encoder, multi-task learning framework that applies the original head than the na"ive multi-task method for both in-domain/zero-shot learning problems. This indicates that we can directly adapt the old formulation as that in the CLIP approach has almost no performance degradation for pure image classification tasks usually directly optimize the parametric classification weights without a need to process text semantics in class names. The CLIP method can be regarded as generating classifier weights through a text encoder and learns the text encoder instead. The text-encoder-based classifier allows sharing between alt-texts as well as modeling their relationships, which enables the ability to tackle any classes.

Although the linear classifier and direct classifier weight parameterization have been common practice in image classification for many years, it is interesting to find that changing the old formulation as that in the CLIP approach has almost no performance degradation for pure image classification problems. This indicates that we can directly adapt the image classification formulation to the cosine classifier and the text encoder parameterization used by CLIP, with almost no loss. This also allows us to further share the text encoder for both class names and alt-texts. Our experiments show that this deep fusion approach performs much better than the naive multi-task method for both in-domain/zero-shot classification and multi-modal retrieval tasks learning (see 3).

Another gap between the image classification and CLIP lies in the different text richness. Class names are usually in short, i.e., one or a few words, and sometimes are even ambiguous and polysemous in referring to specific semantics, for example, “night bird” can represents either “owl” or “nightingale”. On the contrary, alt-texts in CLIP are usually full sentences containing rich information. To further bridge the gap between the image classification and CLIP, we propose a second technique that leverages the knowledge base to enhance the original class names, such as the explanations in dictionaries. In our implementation, knowledge is simply encoded as a prefix/suffix prompt, as illustrated in Fig 1. Although simple, dictionary enhanced method shows to maintain the accuracy for pure image classification problem (see Table 1), while greatly improve the zero-shot and multi-modal retrieval performance as shown in Table 2 and 3. Note the process is just like human beings who learn new words or concepts through both real examples and explanations in dictionaries.

By these techniques, we present a framework that deeply fuses the two important tasks of image classification and contrastive language-image pre-training, dubbed iCLIP. Extensive experiments using different combinations of image classification and image-text pair datasets show that the iCLIP method can take advantages of both the discriminative power of image classification tasks and the zero-shot ability in CLIP-like tasks, and perform significantly better than conducting each task alone or the naive multi-task learning in both the in-domain/zero-shot classification and multi-modal retrieval problems. The iCLIP method also shows that learning a stronger transferable representation than using each of the two tasks alone, verified on a variety of downstream tasks, including ADE20K semantic segmentation [68], LVIS long-tail detection [17], and video action recognition [26], as well as different evaluation settings of few-shot and fine-tuning. Our contributions are summarized as follows:

• We combined two important vision tasks of image classification and contrastive language-image pre-training into a single framework.

• We found that the original image classification formulation can be adapted to CLIP approach with almost no performance degradation. With this finding, we present a deep fusion approach in which the two tasks share the same text encoder and the same classifier type, whose effectiveness is extensively verified on benchmarks.

• We proposed a simple yet effective method to introduce knowledge bases into image classification, addressing the ambiguous and polysemous issue of the originally short image names as well as further bridges the gap between classes and alt-texts. It also provides the first showcase of applying knowledge bases into computer vision problems.

2. Related Work

Supervised visual classification. Classification is almost ubiquitous for visual understanding tasks of various recognition granularity, e.g., image-level classification [12, 20, 28, 33, 49, 52, 58], object-level classification in
object detection [3, 15, 19, 45], pixel-level classification in semantic/instance segmentation [5, 35, 63], and video-level action classification [4, 13, 34, 43, 54]. In these tasks, the data is manually annotated to a fixed set of classes, e.g., the 1,000-class ImageNet-1K dataset [8], the 80-class COCO detection dataset [31], the 150-class ADE20K segmentation dataset [68], etc. Among these classification tasks, the image-level classification is particularly important, which has greatly advances the success of deep learning in computer vision, thanks to its high quality and transferable discriminative representations.

The supervised visual classification is generally performed as a K-way classification problem without considering the text semantics of the class names. The most common classifier is the linear classifier, where the classifier vector of each category is parameterized as model weights and is directly learnt through optimization [28].

**Contrastive language-image pre-training.** Pioneered by CLIP [44] and Align [24], the contrastive language-image pre-training is now attracting more and more attention due to its strong zero-shot transfer capacity. These methods learn a network to pair an image and its associated alt-text, in which the image-text pairs are crawled from the Internet. With web-scale alt-text, it is possible to cover almost all classes, and these methods do show to perform very well for zero-shot recognition. In their frameworks, the images and texts are embedded using two separate encoders, and the output representations of the images and alt-texts are contrasted according to the positive and negative pairs.

While prior to CLIP and Align, there have been a few early works leveraging alt-text or text encoders for image recognition [10, 14, 16, 25, 41, 46, 67]. More follow-up works appeared after CLIP and Align, including Filip [62], De-Clip [30], BASIC [42], LiT [66], LiMoE [39], TCL [60], and so on. A drawback of these methods is that the image-text pairs are usually noisy without human labeling, leading to the learned representations are not conceptual compact, lacking strong discrimination ability.

**Introducing knowledge into AI systems.** Our approach is also related to the expert systems in 1980s which heavily rely on a knowledge base for reasoning [23]. Recently, in natural language process, there also emerges boosting large-scale pretrained models by making use of encyclopedic [1, 55] and commonsense knowledge [50]. However, in computer vision, the knowledge bases is not well explored. We hope our findings can encourage more attention to incorporate human knowledge into current vision systems.

**Combination of representation learning.** Regarding individual strengths of different representation learning approaches, there have been several works trying to combine different representation learning approaches so as to take advantages of individuals’ strength. For example, SLIP [37] combines CLIP learning with a self-supervised contrastive learning approach. CoCa [65] combines the CLIP target with an image caption task, in hope to perform well for both understanding and generation problems. MaskCLIP [11] combines CLIP with masked image modeling based self-supervised learning. In contrast, our work also aims to effectively combine different representation learning approaches so as to take both advantages, specifically, the image classification and CLIP.

**Relationship to UniCL [61]** Concurrent to our work, there is another work named UniCL [61] which also combines image classification with language-image pre-training. We hope the consistent knowledge will help the community in learning more powerful representations. Also note that there are two main differences comparing our framework to the UniCL framework [61]: 1) We involve all negative classifiers in training the supervised classification, while UniCL only involve negatives in a same batch. To make feasible all negative classifiers, we propose a GPU-distributed implementation that distributes the classifiers evenly into different GPUs. Our implementations show to have better in-domain accuracy compared to UniCL when the category number is as large as tens of thousands (76.3% vs. 70.5% as shown in Tab. 4). 2) We introduce a new dictionary enhanced approach to convert the class names with rich semantical text, which shows to be very beneficial for zero-shot image classification and multi-modal retrieval (see Tab. 2).

### 3. Method

In this section, we first review existing methods on image classification and contrastive language-image pre-training tasks. Then, we propose a unified framework to bridge the two tasks in a deep fusion fashion. Finally, we introduce dictionary-enhanced category descriptions to further align the two tasks on input label space.

**3.1. Preliminaries**

**Image Classification.** Given a set of <image, category label> pairs, i.e., \(\mathcal{D}^c = \{(I_i, C_i)\}_{i=1}^{\mathcal{D}^c}\), image classification task targets to predict the category label of a given image, through a visual encoder \(f_v\), and a parametric category classifier \(h_{c}\), illustrated in Fig. 2 (b). The parameters of \(h_{c}\) is a matrix \(W \in \mathbb{R}^{N \times H}\), where \(N\) is the number of categories and \(H\) is the dimension of visual embeddings. The visual encoder \(f_v\) transforms each raw image \(I_i\) to an embedding \(v_i = f_v(I_i)\), while the classifier \(h_{c}\) predicts the distribution \(P_i \in \mathbb{R}^{N}\) over all pre-defined categories via an inner product between \(W\) and \(v_i\), i.e., \(P_i = W \cdot v_i\) (bias term is omitted for simplicity). Finally, a cross entropy is applied on \(P_i\) and \(C_i\) to calculate training loss, which is formulated as:

\[
\mathcal{L} = \frac{-1}{|\mathcal{D}^c|} \sum_{(I_i, C_i) \in \mathcal{D}^c} \log \frac{\exp(W_{C_i} \cdot v_i)}{\sum_{j=1}^{N} \exp(W_j \cdot v_i)},
\]


where $W_j$ is the parametric weight of $j$-th category.

**Contrastive Language-Image Pre-training.** Given a set of $\langle\text{image}, \text{alt-text}\rangle$ pairs, i.e., $D^a = \{(I_i, T^a_i)\}_{i=1}^{|D^a|}$, contrastive language-image pre-training targets to close the distances between paired image and text while enlarging those of unpaired ones, through a visual encoder $f_I$ and a text encoder $f_T$, shown in Fig. 2 (a). They transform the image $I_i$ and the alt-text $T^a_i$ to feature embeddings $v_i$ and $s_i$, respectively. A contrastive loss function is applied to shrink the cosine distance of $v_i$ and $s_i$, which is defined as:

$$
\mathcal{L} = \frac{-1}{|D^a|} \sum_{(I_i, T^a_i) \in D^a} \log \frac{\exp \left( \cos \left( f_T(T^a_i), v_i \right) / \tau \right)}{\sum_{T^a_j \in T^a} \exp \left( \cos \left( f_T(T^a_i), v_j \right) / \tau \right)},
$$

where $\cos(\cdot, \cdot)$ represents the cosine similarity between two embeddings, $T^a_i$ is all the alt-texts in a batch including one positive paired alt-text and $|T^a| - 1$ negative ones, and $\tau$ is a temperature hyper-parameter to scale the similarities.

**Task differences.** Comparing the formations of image classification and language-image pre-training, we can draw three main difference between them. 1) **Training loss functions.** Classification commonly adopts a cross-entropy loss on inner-product similarity, while image-text learning uses InfoNCE loss on cosine similarity. 2) **Classifier types.** Classification adopts a parametric category classifier, while image-text learning uses a text encoder. 3) **Label granularity.** Category names in classification are usually very short, i.e., one or few words, while the captions in image-text pre-training are full sentences containing rich semantics.

### 3.2. Bridge Image Classification and Contrastive Language-Image Pre-training

To bridge image classification and image-text alignment, we introduce three adaptations to align their training losses, unify the classifier types, and close the label granularity gap. The overall adaption is visualized in Fig. 3.

**Classification with Text Encoder.** As formulated in Eq. (1), image classification commonly adopts a cross-entropy loss on top of the inner-product similarity between the visual embedding $v_i$ and the parametric classifier $h_c$. This formulation is not in line with the InfoNCE loss in Eq. (2), leading to a misalignment between the two paradigms. To address this issue, we adopt a cosine similarity for image classification, instead of the original inner-product similarity in Eq. (1), which formulates a cosine classifier as:

$$
\mathcal{L} = \frac{-1}{|D^e|} \sum_{(I_i, C_i) \in D^e} \log \frac{\exp \left( \cos \left( W_{C_i}, v_i \right) / \tau \right)}{\sum_{j=1}^{N} \exp \left( \cos \left( W_j, v_i \right) / \tau \right)},
$$

Cosine similarity is a common practice in metric learning [40]. It can smoothly align the supervised image classification with the cross-modal contrastive pre-training in terms of learning objective function, i.e., Eq. (2). Moreover, our experiments demonstrate that this cosine classifier performs on par with the traditional linear classifier (see Tab. 1).

The cosine classifier aligns the training losses of two tasks. However, the annotations, i.e., category labels and captions, are modeled separately by the parametric category classifier $h_c$ and the text encoder $f_T$. As analyzed in Sec. 4.3, shallowly combining the two tasks with a shared...
visual encoder \( f_v \) and two separate task heads does not fully take advantage of the gold annotations in image classification and rich concepts in textual captions, resulting in a sub-optimal solution with limited transferring capacity.

To tackle this issue, we take label semantics into consideration and propose to utilize the text encoder \( f_t \) as a meta classifier for image classification. Formally, we replace the label index \( C_i \) with its class name \( M_i \), and generate the classifier weight \( W \) on-the-fly through the text encoder \( f_t \), which is shared with image-text pre-training. The new formulation is represented as:

\[
L = -\frac{1}{|D^c|} \sum_{(i, M_i) \in D^c} \log \frac{\exp (\cos (f_t(M_i), v_i) / \tau)}{\sum_{j=1}^{|D|} \exp (\cos (f_t(M_j), v_i) / \tau)}. \tag{4}
\]

In this way, the text encoder \( f_t \) is not only used to extract semantics from gold category labels, but also capture textual information from image captions. Both the visual and textual encoders are shared across the two tasks, leading to a deep fusion of the two tasks.

Classification with Dictionary Enhancement. The cosine classifier with text encoder as a meta network has largely unify the two tasks in model training. In this step, we further align them on input label granularity, reducing the disparity between label names (one or few words) and image captions (a complete sentence). Our proposal is to integrate external knowledge into label names. More specifically, for each label names, we introduce detailed descriptions from its corresponding synset in the dictionary WordNet [36] as the external knowledge and create a pseudo sentence as label for each categories. We combine the original class names and their dictionary descriptions to form the enhanced texts as the input to the text encoder. Also, we add a prompt to make the sentence more fluent. The final dictionary-enhanced description for each category is formed as:

\[
\mathcal{T}^c = \text{A photo of a } \{\text{NAME}\}_{C_i}, \{\text{DESCRIPTION}\}_{C_i}. \tag{5}
\]

Such dictionary-enhanced descriptions have similar label granularity to alt-text, and thus further bring image classification closer to image-text alignment. Moreover, the description introduces more details of each category, being capable of reducing potential misconception. For example, the class “night bird” actually includes several kinds of birds, like owl, nightingale, etc. Such a category name cannot allow the model to learn precise representations due to the blurry concepts. If we augment the category with more external knowledge, such as “a photo of a night bird, any bird associated with night: owl, nightingale, nighthawk”, it will help the model learn discriminative representation on distinguishing different concepts (e.g., bird species).

A Unified Framework. The above three steps adapt image classification to image-text alignment from the perspective of training loss, classifier type and annotation granularity, respectively. Towards the final unification, we propose a new framework dubbed iCLIP, as presented in Fig. 2 (c), which bridges Image Classification and Image-Text Alignment with a unified contrastive learning loss formulated as:

\[
\mathcal{L} = -\frac{1}{|D|} \sum_{(i, T_i) \in D} \log \frac{\exp (\cos (f_t(T_i), v_i) / \tau)}{\sum_{j \in \mathcal{T}} \exp (\cos (f_t(T_j), v_i) / \tau)}. \tag{6}
\]

where \( D \) is a set consisting of the image classification data \( D^c \) and the image-text alignment data \( D^a \), i.e., \( D = \{D^c, D^a\} \), while \( \mathcal{T} \) indicates a combination of \( \mathcal{T}^c \) and \( \mathcal{T}^a \), i.e., \( \mathcal{T} = \{\mathcal{T}^c, \mathcal{T}^a\} \). Text label \( T_i \) is either an image caption \( T_i^c \) or a dictionary-enhanced description \( T_i^a \) sampled from \( \mathcal{T} \). It is worth noting that, with this unified framework, both the text encoder \( f_t \) and the visual encoder \( f_v \) are shared across the two tasks, achieving a deep fusion. The proposed unified framework is able to leverage any combination of tag-labeled and caption-labeled image datasets for pre-training. This combination allows the model to learn more discriminative representation, while capturing more visual concepts from the textual description. On the other hand, our iCLIP method is efficient.

Distributed Implementation. In our iCLIP framework, the text embedding of each category is generated by the shared text encoder on-the-fly. This computation is affordable when the number of categories \( N \) is not large. However, it will become infeasible if category number scales up to be large, such as 22k categories in ImageNet-22K [8]. To make the iCLIP framework feasible for large-category classification data in practice, we adopt a distribution implementation strategy [6]. Specifically, we distribute all the enhanced class names evenly over \( G \) GPUs in forward, and gather the embeddings from each gpu for similarity calculation, reducing the computation cost and saves memory consumption by the text encoder to \( 1/G \).

<table>
<thead>
<tr>
<th>#</th>
<th>Cosine Loss</th>
<th>Text-enc. as Enhanced Classifier</th>
<th>Enhanced classes</th>
<th>IN-1K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.9</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.5</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.2</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.4</td>
</tr>
</tbody>
</table>

4. Experiment

We verify the effectiveness of the proposed iCLIP framework through the comparisons to single-task baselines and a naïve multi-task learning baseline. The comparisons are conducted in three settings covering different scales of pre-training data. In evaluation, we assess the models on different tasks, including in-domain classification, zero-shot classification, multi-modal retrieval, and downstream tasks.
4.1. Experimental Setup

**Pre-training data and settings.** We consider three different scales of dataset combination for model pre-training.

- **ImageNet-1K [8] and GCC-3M [48].** In this setting, we use ImageNet-1K as the classification data while GCC-3M as the image-text data. We adopt a Swin-T [33] initialized with MoBY [59] as the visual encoder, while for the textual encoder, we use a pretrained RoBERTa-B [32]. We sample half number of images from each dataset in a mini-batch and train the models with a batch size of $128 \times 8$ V100 GPUs for 100 epochs. The highest learning rate is $2e-4$ with a cosine learning rate schedule and 5 epochs warm-up. Weight decay is set to be 0.01. RandAugment [7] and stochastic depth [21] with a rate of 0.1 are used for visual encoder only.

- **ImageNet-22K [8] and YFCC-14M [53].** We follow UniCL [61] to train all models from scratch with 32 epochs for a fair comparison with it. Swin-T [33] is used as the visual encoder, and a 12-layer transformer with a hidden dimension of 512 same as CLIP [44] is used as the text encoder. A batch size of $512 \times 16$ GPUs is adopted. The highest learning rate is selected from 2e-4 and 8e-4. Other regularization is the same as previous, except for a larger weight decay of 0.05. We also conduct experiments using two variants of this setup for a fair and clean comparison with the methods that use one task alone (IC or CLIP): 1) Excluding the 1,000 ImageNet-1K classes in ImageNet-22K dataset (dubbed IN-21K). This setup variant allows us to evaluate the zero-shot accuracy on ImageNet-1K for different methods; 2) Half images of the ImageNet-21K and YFCC-14M are used, such that the dataset size and training iterations are the same as that used in one single task.

- **ImageNet-22K [8] and Laion-400M [47].** For this large-scale pre-training setting, we adopt a Swin-B initialized with MoBY as the visual encoder and a pre-trained RoBERTa-B as the text encoder. We train iCLIP for 100K iters, with a batch size of $192 \times 64$ V100 GPUs. In each mini batch, we sample 64 images from IN-22K and 128 images from Laion-400M. The model is trained on classification data for around 30 epochs and on image-text data for around 2 epochs equivalently. The highest learning rate is $1e-3$ with a cosine learning rate schedule and a warm-up for 16.7K iters. Weight decay is set to 0.05 and drop depth rate is set to 0.2.

**Evaluation datasets and settings.** During evaluation, we assess the models considering five different settings.

- **Zero-shot classification.** We evaluate the concept coverage and generalization ability of the models on three datasets: 1) ImageNet-1K variants, including IN-1K [8], and IN-Sketch (IN-S) [56]. Top-1 accuracy is reported; 2) the widely-used Kornblith 12-dataset benchmark [27]; 3) 14 datasets used in UniCL [61]. For 2) and 3), averaged accuracy is reported.

- **Zero-shot multi-modal retrieval.** Flickr30K [64] (1K test set) and MSCOCO [31] (5K test set) are used to evaluate the alignment between image and text modalities. We report the Top-1 recall on both image retrieval (IR) and text retrieval (TR).

- **In-domain classification.** ImageNet-1K data is included in some of our pre-training setups, so we conduct in-domain evaluation on ImageNet-1K in these cases. The Top-1 accuracy is reported.

- **Few-shot classification.** Following CLIP [44], we also evaluate the models on few-shot classification task using Kornblith 12-dataset with a frozen visual encoder. Averaged accuracy is reported.

- **Fine-tuning on downstream tasks.** To validate the generalization ability of iCLIP, the models are fine-tuned and compared on semantic segmentation [68], long-tail detection [17], and video action recognition [26]. We report val mIoU, bbox mAP and Top-1 accuracy, respectively. The detailed settings can be found in the supplementary material.

### 4.2. Experiments on IN-1K [8] and CC3M [48]

**Formulation adaptations for image classification.** Tab. 1 ablates the effect of adapting the common image classification to that used in iCLIP, including both cosine loss, the
Table 3. Ablation conducted on IN-1K [8] and GCC-3M [48] combined data. For the models only using IN-1K, we train them for 100 epochs. For the models only using GCC-3M, we train them with the same iterations and batch size as the ones used in IN-1K.

<table>
<thead>
<tr>
<th>Method</th>
<th>12-dataset avg.</th>
<th>ImageNet-related avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IN-1K</td>
<td>IN-S</td>
</tr>
<tr>
<td>1. Sup-only</td>
<td>34.4</td>
<td>32.4</td>
</tr>
<tr>
<td>2. VL-only</td>
<td>40.5</td>
<td>80.6</td>
</tr>
<tr>
<td>3. Native multi-task</td>
<td>37.7</td>
<td>80.5</td>
</tr>
<tr>
<td>4. iCLIP (w/o Desc.)</td>
<td>39.1</td>
<td>80.4</td>
</tr>
</tbody>
</table>

Table 4. Comparison with UniCL. Models are pre-trained from scratch with 32 epochs, following UniCL [61].

<table>
<thead>
<tr>
<th>#</th>
<th>Training Data</th>
<th>Method</th>
<th>IN-1K</th>
<th>14-dataset avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YFCC + IN-21K (half)</td>
<td>UniCL [61]</td>
<td>35.1</td>
<td>45.0</td>
</tr>
<tr>
<td>2</td>
<td>YFCC + IN-21K (half)</td>
<td>iCLIP</td>
<td>35.1</td>
<td>45.0</td>
</tr>
<tr>
<td>3</td>
<td>YFCC + IN-21K</td>
<td>UniCL [61]</td>
<td>40.5</td>
<td>49.1</td>
</tr>
<tr>
<td>4</td>
<td>YFCC + IN-21K</td>
<td>iCLIP</td>
<td>50.9</td>
<td>54.4</td>
</tr>
<tr>
<td>5</td>
<td>YFCC + IN-22K</td>
<td>UniCL [61]</td>
<td>70.5</td>
<td>52.4</td>
</tr>
<tr>
<td>6</td>
<td>YFCC + IN-22K</td>
<td>iCLIP</td>
<td>76.3</td>
<td>55.5</td>
</tr>
</tbody>
</table>

text-encoder-based classifier and enhanced class names using ImageNet-1K dataset. It can be seen that the cosine classification loss gets slightly better performance than the linear one, with a +0.6% gain on IN-1K (see #1 v.s. #2). When further adapting the text-encoder-based classifier (#3) and enhancing class names from dictionaries (#4), it has almost no performance degradation (+0.3% and +0.5% on IN-1K compared to the linear classifier), which allows to further sharing the text encoder with CLIP for tasks unification.

Zero-shot and in-domain classification. With previous adaptions on the image classification formulation, we can further share the text encoder between the two tasks. To ablate the effect of sharing the text encoder, we set a naïve multi-task baseline, that combines image classification and CLIP in a shallow fusion, i.e., simply averaging the loss Eq. (1) and Eq. (2). Each has its own head network, i.e., the fully-connected layer $W$ for Eq. (1) and the text encoder $f_t$ for Eq. (2). The best performances of the two heads are reported in Tab. 3. With a shared text encoder across the two tasks, our iCLIP (w/o Desc.) outperforms the naïve multi-task on Kornblith 12-dataset zero-shot classification by +2.6% in average, while they are comparable on ImageNet-related datasets classification (see #3 v.s. #4). Our iCLIP deeply unifies two tasks, thus better gathering the merits of the two learning protocols. When compared with the supervised softmax classifier baseline, i.e., Eq. (1) Sup-only, and the contrastive image-text pre-training baseline, i.e., Eq. (2) VL-only, our method is slightly worse than Sup-only on IN-1K by 0.4%, while achieves superior performance on other evaluation settings, +6.5% better than VL-only method on 12-dataset zero-shot testing and +9.2% better than Sup-only method on IN-S (see #4 v.s. #1 & #2). Moreover, the dictionary enhancement on class names (#5) can further bring an average of +1.4% improvements on Kornblith 12-dataset, revealing the increased discriminative representation for ambiguous concepts.

4.3. Experiments on IN-22K [8] and YFCC14M [53]

Effects of the unified framework. Here, we further ablate the effect of the unified formulation for deep fusion of the two tasks. In #2, #4 and #6 of Tab. 2, we show the results of our unified framework under three different dataset combination setups. Compared with the CLIP baseline (#1), our iCLIP (#2) earns +8.3% gains on IN-1K zero-shot classification and also +9.1% improvements when evaluated on the 14-dataset. In addition, our iCLIP is better than the CLIP baseline on most cross-modal retrieval benchmarks, while only using half of visual-language data in pre-training.

Effects of dictionary enhancement. Furthermore, we dissect the model to study the contributions of dictionary-enhanced category description. From Tab. 2, we can see that enhancing each class names with informative description from the dictionary brings consistent improvements on both zero-shot classification and zero-shot retrieval under three dataset combination setups (see #3, #5 and #7). In particular, when pre-trained with half images of YFCC-14M and IN-21K (#3), the integrated knowledge contributes +6.5% improvements on IN-1K zero-shot classification, which makes our iCLIP reach 45.9%, being +5.4% better than UniCL method [61] with full images of YFCC-14M and IN-21K (see #3 in Tab. 4). More importantly, the enhanced class names is beneficial to cross-modal retrieval. For example, for image-to-text search, the dictionary-enhanced description can bring 10.7% and 6.8% top-1 recall gains on Flickr30K [64] and MSCOCO [31] respectively, as reported in row 3 of Tab. 2.

Comparison with UniCL [61]. Tab. 4 summaries our comparison to UniCL. The same as UniCL, we evaluate our models on IN-1K and 14 datasets. Under three different dataset combination setups, our iCLIP surpasses UniCL by at least +5% on IN-1K image classification, while reaching 55.5% averaged accuracy on 14 datasets (#6), being +3.1% better than UniCL (#5).

4.4. Experiments on IN-22K and Laion-400M [47]

Zero-shot and in-domain classification. Tab. 5 presents a large scale experiment using the publicly accessible large-scale data: Laion-400M [47] and IN-22K [8]. For Sup-only, i.e. Eq. (1), we use the released version from Swin [33], which is trained on IN-22K for 90 epochs. For VL-only, i.e. Eq. (2), we pre-train it on Laion-400M with a similar image numbers (#im). Our method is comparable to Sup only on IN-1K, while it gets +17.8% and +8.3% better results than the two baselines on IN-S, demonstrating its robustness to natural distribution shifts. Our iCLIP surpasses
Table 5. Ablation study on IN-22K [8] and Laion-400M [47]. We evaluate the models on ImageNet datasets (IN-1K [8] and IN-S [56]) and zero-shot evaluation on the Kornblith 12-dataset benchmark [27]. Few-shot learning on Kornblith 12-dataset and the fine-tuning on three downstream tasks are conducted to evaluate the transfer capability of iCLIP. ‡ denotes for our reproduction using released checkpoints.

<table>
<thead>
<tr>
<th>Method</th>
<th>Visual encoder</th>
<th>Pre-train</th>
<th>ImageNet-related 12-dataset avg.</th>
<th>downstream tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arch.</td>
<td>length (#im.)</td>
<td>IN-1K</td>
<td>IN-S</td>
</tr>
<tr>
<td>CLIP [44]</td>
<td>ViT-B/16</td>
<td>400 × 32 eps</td>
<td>68.6</td>
<td>46.6‡</td>
</tr>
<tr>
<td>OpenCLIP [22]</td>
<td>ViT-B/16</td>
<td>400 × 32 eps</td>
<td>67.1</td>
<td>52.4‡</td>
</tr>
<tr>
<td>Sup-only</td>
<td>Swin-Base</td>
<td>14M × 90 eps</td>
<td>82.6</td>
<td>42.0</td>
</tr>
<tr>
<td>VL-only</td>
<td>Swin-Base</td>
<td>400M × 3 eps</td>
<td>61.1</td>
<td>51.5</td>
</tr>
<tr>
<td>iCLIP</td>
<td>Swin-Base</td>
<td>400M × 2 eps + 14M × 30 eps</td>
<td>82.9</td>
<td>59.8</td>
</tr>
</tbody>
</table>

When only given one example per class, by utilizing text encoder as the classifier, our iCLIP achieve 73.9% on 12-dataset in average, surpassing the original CLIP model by +29.5%. Such one-shot recognition gets +3.3% gains over the zero-shot baseline (⋆ v.s. ⋅ ⋅), demonstrating good few-shot transfer ability. When using 16 examples per class, our model still performs superior to CLIP by 4.1%. Compared to supervised-only model and visual-linguistic only model, our unified contrastive learning pretrained model obtains +24.6% and +6.1% better accuracy under one-shot learning setting. Such advantages are kept to 16-shot with +2.7% and +5.0% gains (⋅ ⋅ and ⋅ ⋅).

Fine-tuning on Downstream Tasks We also study the generalization capability of our pre-trained models on downstream tasks, including semantic segmentation, object detection and video recognition. As shown in Tab. 5, compared to Sup-only, our iCLIP surpasses it by +0.5%, +2.0%, +0.4% on the three downstream tasks, respectively. We also earn +0.6%, +1.3%, +0.8% gains over VL-only baseline. These results reveal that our unified method could learn general visual representations.

5. Conclusion

In this paper, we propose a unified framework dubbed iCLIP to bridge image classification and language-image pre-training. It naturally forces the cross-modal feature learning in a unified space, where the two tasks share the same visual and textual encoders. Extensive experiments demonstrate that iCLIP is effective, and can be generalized to different visual recognition scenarios, including zero-shot, few-shot, and fully-supervised fine-tuning.

Limitations. One limitation of iCLIP is that, despite its competitive performance, the model still relies on human labeled classification data that is not scalable. Besides, our model currently only adopts median-size parameters, which can not fully validate the generation ability to large-scale models. We are interested in exploring this in future work.
References


[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics hu-

[27] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2656–2666, 2019. 6, 8


[38] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. 2020. 2


[53] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and
Li-Jia Li. Yfcc100m: The new data in multimedia research. 
Commun. ACM, 59(2):64–73, jan 2016. 6, 7


[56] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In Advances in Neural Information Processing Systems, pages 10506–10518, 2019. 6, 8


[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78, 2014. 6, 7


2786