Multi-Level Cross-Modal Alignment for Image Clustering

Liping Qiu*, Qin Zhang*, Xiaojun Chen†, Shaotian Cai
Shenzhen University, Shenzhen, China
qiuliping2021@email.szu.edu.cn, {qinzhang, xjchen}@szu.edu.cn, cai.st@foxmail.com

Abstract

Recently, the cross-modal pretraining model has been employed to produce meaningful pseudo-labels to supervise the training of an image clustering model. However, numerous erroneous alignments in a cross-modal pretraining model could produce poor-quality pseudo labels and degrade clustering performance. To solve the aforementioned issue, we propose a novel Multi-Level Cross-Modal Alignment method to improve the alignments in a cross-modal pretraining model for downstream tasks, by building a smaller but better semantic space and aligning the images and texts in three levels, i.e., instance-level, prototype-level, and semantic-level. Theoretical results show that our proposed method converges, and suggests effective means to reduce the expected clustering risk of our method. Experimental results on five benchmark datasets clearly show the superiority of our new method.

Introduction

Image clustering which groups images into different clusters without labels is an essential task in unsupervised learning. Many methods are proposed to utilize the large-scale pre-training models such as Resnet (He et al. 2016) or ViT (Dosovitskiy et al. 2020) to extract high-quality representations for image clustering (Ji, Vedaldi, and Henriques 2019; Li et al. 2021; Zhong et al. 2021; Wu et al. 2019; Van Gansbeke et al. 2020; Dang et al. 2021). Then, to unsupervised classification models, multiple indirect loss functions are used (e.g. sample relations (Chang et al. 2017a), invariant information (Ji, Vedaldi, and Henriques 2019; Li et al. 2021), mutual information (Wu et al. 2019) and entropy (Huang, Gong, and Zhu 2020; Van Gansbeke et al. 2020; Li et al. 2021). However, as pointed out in (Cai et al. 2023), the aforementioned techniques have difficulties in handling examples that are semantically different but visually comparable by focusing only on images.

Recently, many vision-language pre-training (VLP) models have been developed to align images and texts into a unified semantic space (Li et al. 2019; Chen et al. 2020; Ramesh et al. 2021; Li et al. 2020; Radford et al. 2021; Jia et al. 2021). To utilize the VLP models for image clustering, Cai et al. (Cai et al. 2023) proposed to use CLIP (Radford et al. 2021) to produce meaningful pseudo-labels and achieved significant improvements on a wide range of datasets in comparison with conventional image clustering methods. Li et al. (Li, Savarese, and Hoi 2022) also used CLIP for a zero-shot image classification task. The success of these methods suggests a promising direction for image clustering. However, as depicted in Figure 1, there are instances where the alignments between images and texts in CLIP may be incorrect for downstream tasks, resulting in subpar pseudo labels and poor clustering performance. SIC (Cai et al. 2023) simply uses CLIP to obtain the embeddings of images and texts and cannot deal with incorrect alignments. While MUST (Li, Savarese, and Hoi 2022) strives to optimize the image encoder in CLIP, its efficiency is hampered by using the pretraining task to update the image encoder, resulting in a slow process.

To address the above problem, we propose a novel method, namely Multi-Level Cross-Modal Alignment (MCA), an efficient way to improve the alignments between images and texts in CLIP for clustering tasks. In general, our main contributions are as follows:

- We propose to use the hierarchical structure in WordNet (see Figure 1) to filter irrelevant words and construct a
smaller but better semantic space, thus reducing the affection of unrelated nouns for clustering. Our experimental findings demonstrate that it can reduce the number of words by up to 60% and significantly enhance clustering performance compared to SIC.

- We propose to optimize both image and text embeddings for downstream tasks, by aligning the images and texts at three levels, i.e., instance-level, prototype-level, and semantic-level. Our proposed method can better fix the incorrect alignments in CLIP for downstream tasks when compared to SIC and MUST.

- Theoretical findings demonstrate that our proposed method converges at a sublinear rate and offers effective strategies for lowering the expected clustering risk of our method. These findings will provide valuable guidance for the design of new image clustering methods.

- Experimental results on five benchmark datasets clearly show the superiority of our new method, especially when dealing with complex clusters.

Related Work

The early deep clustering methods simply combine representation learning and shallow clustering (Xie, Girshick, and Farhadi 2016; Yang et al. 2017; Tian, Zhou, and Guan 2017; Shaham and Stanton 2018). With the rapid development of the pre-training paradigm, many methods employ large-scale pre-training models such as Resnet (He et al. 2016) or ViT (Dosovitskiy et al. 2020) to extract high-quality representations and train a classification model, by maximizing the consistency between each image and its augmentation-s/neighbors (Ji, Vedaldi, and Henriques 2019; Li et al. 2021; Zhong et al. 2021; Wu et al. 2019; Van Gansbeke et al. 2020; Zhong et al. 2021; Dang et al. 2021), or generating pseudo-labels (Wu et al. 2019; Van Gansbeke et al. 2020). However, as pointed out in (Cai et al. 2023), it is challenging for the aforementioned techniques to handle examples that are semantically different but visually comparable by only accessing visual information in images.

Cross-modal clustering has made significant progress in recent years, which usually learns a shared subspace such that the mutual agreement between multiple modalities is maximized, by Canonical Correlation Analysis (CCA) (Gao et al. 2020) or mutual information optimization (Mao et al. 2021). However, these methods require image-text pairs as input, which may be cost-intensive to collect in real applications.

Recently, vision-language pre-training (VLP) models that align multi-modal data in common feature space by different pre-training tasks have been proposed. For example, VisuAlBert (Li et al. 2019), UNITER (Chen et al. 2020) and DALL-E (Ramesh et al. 2021) use language-based training strategies, including mask LM (Language Modeling) such as Masked Language/Region Modeling, or autoregressive LM such as image caption and text-grounded image generation. CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) utilize cross-modal contrastive learning to align the visual and textual information into a unified semantic space.

Since VLP captures the relationships among images and texts (low-level semantics), it is natural to utilize VLP models to compensate for the semantic information for better image clustering. Cai et al. (Cai et al. 2023) proposed to use CLIP (Radford et al. 2021) to generate meaningful pseudo-labels for image clustering. Li et al. (Li, Savarese, and Hoi 2022) also proposed to use CLIP for zero-shot image classification tasks.

Notation and Problem Definition

Suppose we have an image dataset \( X = \{x_1, x_2, \ldots, x_n\} \) with \( n \) instances sampled i.i.d. from input space \( D \), we can obtain the embeddings of these images as \( U = \{u_1, u_2, \ldots, u_n\} \) where \( u_i = e_I(x_i) \in \mathbb{R}^{d \times 1} \) is obtained via the image encoder \( e_I(.) \) of CLIP, where \( d \) is the embedding dimension. To capture the semantic meaning of these images, we first introduce a noun vocabulary \( T = \{t_1, t_2, \ldots, t_m\} \) that includes \( m \) noun phrases sampled from WordNet (Miller 1995). Then we can obtain the embeddings of these \( m \) words as \( V = \{v_1, v_2, \ldots, v_m\} \) where \( v_i = e_T(s_i) \in \mathbb{R}^{d \times 1} \), \( s_i \) is a sentence like “A photo of a \{t_i\}” and \( e_T(.) \) is the text encoder of CLIP. Let \( c \) be the number of categories, our goal is to group the images in \( X \) into \( c \) clusters with the help of CLIP. Let \( f_I(e_I(X); \phi) \) denotes the image classification network with parameters \( \phi \) that maps an image \( x_i \) into a soft cluster assignment probability vector \( q_i \in \mathbb{R}^{c \times 1} \), and \( f_S(e_T(T); \theta) \) denotes the text classification network with parameters \( \theta \) that maps a word \( t_i \) into a soft cluster assignment probability vector \( p_i \in \mathbb{R}^{c \times 1} \). Notably, \( e_I \) and \( e_T \) in CLIP are kept frozen during the training process.

Method

In this paper, we propose our method shown in Figure 2. This new method mainly consists of three components: 1) Semantic space construction builds a proper semantic space \( T \), 2) Image consistency learning performs the consistency learning in image space, and 3) Multi-level cross-modal alignment aligns images and texts at three distinct levels: instance-level, prototype-level, and semantic-level.

Semantic Space Construction

Constructing a proper semantic space \( T \) from WordNet (Miller 1995) such that the images can be well represented by the words in \( T \) is very important for image clustering, because too small \( T \) may lose important relevant words but too large \( T \) may contain too many irrelevant noisy words. In this step, we first build a candidate semantic space \( \mathcal{W} \) as 82,000 nouns in the WordNet dataset (Miller 1995). Considering an image dataset usually makes up a small part of semantics, we propose a two-step filtering strategy to construct a proper semantic space for an image dataset: 1) Uniqueness-based filtering selects \( \gamma \) nearest words for each of \( c \) image cluster centers obtained by \( \rho \) means of the most unique nouns whose uniqueness scores (Cai et al. 2023) are greater than a given hyperparameter \( \rho_u \). 2) Hierarchy-based filtering employs the hierarchical structure in WordNet to further filter \( \mathcal{W}_c \) to form
the final semantic space \( \mathcal{T} \). Let \( \mathcal{T} = \emptyset \). Given an image \( x_i \), we find its nearest noun \( w_i \in \mathcal{W}_c \) and search its hierarchical structure from WordNet (Miller 1995) to form a hierarchical semantic tree. In general, the words in the lower layers provide more fine-grained information for distinguishing the images, while the words in the higher layers may be useless for clustering. Figure 1 shows, for example, “mammal” is the common parent of “dog” and “cat” and cannot distinguish the images in “dog” or “cat”. Therefore, we propose a novel method for this task, which is shown in Figure 2. Specifically, our method employs a three-level alignment approach. Firstly, at the **Instance-level Alignment**, each image is aligned with its neighboring texts. Secondly, the **Prototype-level Alignment** aligns each image prototype with its nearest text prototype. Lastly, at the **Semantic-level Alignment**, each image is aligned with its neighboring texts in the semantic space. The detailed descriptions of these three alignment processes are provided below.

**Instance-level alignment**: Given an image \( x_i \) and its neighboring texts \( \mathcal{N}_{k^i}(x_i) \), we propose the following contrastive loss function to facilitate the alignment:

\[
L_{ia} = -\frac{1}{n} \sum_{i=1}^{n} \log q_i^T \frac{1}{\tau} \sum_{j=rn(\mathcal{N}_{k^i}(x_i))} \exp(q_i^T p_j / \tau_{ia})
\]

where \( \mathcal{N}_{k^i}(x_i) \) contains \( k^i \) nearest images of \( x_i \) and \( rn(\mathcal{N}_{k^i}(x_i)) \) randomly selects a sample from \( \mathcal{N}_{k^i}(x_i) \). The second item is the popular negative entropy loss for preventing trivial solutions that most samples belong to a small proportion of clusters, where \( \bar{q}_i = \frac{\sum_{l=1}^{c} q_{i,l}}{\tau_{ia}} \) is the average cluster assignment. \( \eta \) is a trade-off parameter.

**Prototype-level alignment**: Instancel-level alignment may be affected by noisy neighborhood relationships, so we further propose to align images and texts at prototype level which is more robust to noisy texts. We first compute an image prototype set \( \mathcal{H}^I \), where \( \mathbf{h}^I_i \in \mathcal{H}^I \) is computed as \( \mathbf{h}^I_i = \frac{1}{|q_i|} \sum_{l=1}^{c} q_{i,l} \mathbf{u}_l \). Then, for each image prototype \( \mathbf{h}^I_i \in \mathcal{H}^I \), we can identify the word in \( \mathcal{T} \) that is closest to \( \mathbf{h}^I_i \) and finally construct a prototype set \( \mathcal{H}^S \). To further improve the prototypes in \( \mathcal{H}^S \), we find \( k_p \) nearest neighborhoods for each \( \mathbf{h}^S_i \in \mathcal{H}^S \) to compute the prototype of these neighborhoods and replace \( \mathbf{h}^S_i \) to update \( \mathcal{H}^S \).

Finally, to align images and texts at the prototype level,
we propose the following loss function:

$$
\mathcal{L}_{pa} = \frac{1}{c} \sum_{j=1}^{c} \log \left( \frac{\exp(f_j(h^j, \phi)^T f_S(h^j, \theta) / \tau_{pa})}{\sum_{l=1, l \neq j}^{c} \exp(f_j(h^j, \phi)^T f_S(h^j, \theta) / \tau_{pa})} \right)
$$

(3)

where $\tau_{pa}$ is a temperature parameter.

**Semantic-level alignment:** Given an image $x_i$ and its neighboring texts $\mathcal{N}_{k_s}(x_i)$, we let $v_j = e_T(s_j)$ be the the embeddings of $t_j \in \mathcal{N}_{k_s}(x_i)$ and $s_j$ is a sentence like “A photo of a $\{t_j\}$”, and $p_j = f_S(v_j, \theta)$. It makes sense that the neighboring texts of an image can help determine the cluster assignment of this image. Note that the alignment relationships between images and texts may vary in different downstream tasks, we propose to use the attention mechanism (Vaswani et al. 2017) to quantify the correlations between an image and its neighboring texts. Specifically, we compute $p'_i$ as the weighted combination of the neighboring texts’ assignments as:

$$
p'_i = f_A(u_i, V^S, P^S; W^I, W^S) = \sum_{j \in \mathcal{N}_{k_s}(x_i)} \text{softmax}((W^I u_i)^T W^S v_j) p_j
$$

(4)

where $f_A(\cdot; W^I, W^S)$ is the attention network to quantify the correlations between an image and its neighboring texts and $W^I, W^S \in \mathbb{R}^{d \times d}$ are two parameter matrices. Then we use the argmax operation to generate one-hot pseudo-label for $x_i$ as:

$$
q'_i = \text{one-hot}(c, \text{argmax}_l p'_il)
$$

(5)

where one-hot($c, l$) generates a $c$-bit one-hot vector with the $l$-th element as 1.

Here, $q'_i$ can be considered as the semantic cluster assignment of $x_i$. Therefore, we perform alignment for each image $x_i$ by aligning the semantic cluster assignment $q'_i$ to the image cluster assignment $q_i$ with the following loss function:

$$
\mathcal{L}_{sa} = \frac{1}{n} \sum_{i=1}^{n} CE(q_i, q'_i)
$$

(6)

where $CE(\cdot)$ is the cross entropy function.

**The overall alignment loss function is:**

$$
\mathcal{L}_A(f_1(e_T(\mathcal{X}); \phi), f_S(e_T(T); \theta), f_A(\cdot; W^I, W^S)) = \mathcal{L}_{sa} + \lambda_{pa} \mathcal{L}_{pa} + \lambda_{sa} \mathcal{L}_{sa}
$$

(7)

where $\lambda_{pa}$ and $\lambda_{sa}$ are two trade-off parameters.

**The Overall Objective**

$$
\mathcal{L}_A(f_1(e_T(\mathcal{X}); \phi), f_S(e_T(T); \theta), f_A(\cdot; W^I, W^S)) = \mathcal{L}(g(S; \varphi)) + \lambda_{a} \mathcal{L}(g(S; \varphi))
$$

(8)

where $\lambda_{a}$ is a trade-off parameter.

**Theoretical Analysis**

In this part, we first analyze the convergence of our proposed method and then analyze its expected clustering risk. We first introduce the following assumptions: **Image Neighborhood Consistency Bound:** $\forall x_i \in \mathcal{X}, x_j \in \mathcal{N}_{k_s}(x_i)$, $q'_i q_j \in [\mu_1, 1]$. **Cross-modal Neighborhood Consistency Bound:** $\forall x_i \in \mathcal{X}, t_j \in \mathcal{N}_{k_s}(x_i), q'_i q_j \in [\mu_C, 1]$. **Image Prediction Confidence Bound:** $\forall x_i \in \mathcal{X}$, $\|q_i\|_\infty \leq \mu_y$. **Image Neighborhood Imbalance Bound:** $\forall x_i \in \mathcal{X}$, $x_i$ is in at most $k'_I$ samples’ (in $\mathcal{X}$) nearest neighborhoods.

We first give the following theorem demonstrating that the optimization algorithm theoretically converges to the local optima in a sublinear speed.

Suppose that $g(S; \varphi)$ is twice differential with bounded gradients and Hessians, and $\mathcal{L}(g(S; \varphi))$ has $L$-Lipschitz continuous gradient. Suppose that the learning rate $\eta_{\varphi}$ satisfies $\eta_{\varphi} = \min \{1, \frac{C}{\sqrt{\tau}} \}$ for some $C > 0$, such that $\frac{C}{\sqrt{\tau}} \geq L$. Then our proposed method can achieve

$$
\min_{0 \leq \epsilon \leq \tau} \mathbb{E} \left[ ||\nabla \mathcal{L}(g(S); \varphi^{(t)})||^2 \right] \leq \epsilon \in O(1/\epsilon^2)
$$

where $\epsilon$ is a very small positive real number.

Next, we analyze the ability of our method to achieve cluster performance on unseen data. Let $\mathcal{L}_{\tau}(g)$ be the empirical clustering risk of MCA and its expectation can be denoted as $\mathcal{L}(g)$. The family of $g$ is defined as $\mathcal{G}$. Then we can obtain the following theorem by analyzing the generalization bound of our proposed method.

Suppose $f_1(\cdot; \phi)$ is Lipschitz smooth with constant $L_I$ and $\|u\|_\infty \leq M_u$. Suppose $\beta \frac{1}{n} \sum_{i=1}^{n} q_i u_i = f_1(h^I, \phi)^T f_S(h^S, \theta)$ is $L_{I,S}$-Lipschitz continuous, where $h^I$ and $h^S$ are computed according to the method in Section $4$. For any $0 < \delta < 1$, we can guarantee that with a probability of at least $1 - \delta$ for any $g \in \mathcal{G}$, the following inequality holds.

$$
\mathcal{L}(g) \leq \hat{\mathcal{L}}_{\tau}(g) + \hat{\epsilon}_1 \frac{1}{\sqrt{n}} + \hat{\epsilon}_2 \frac{1}{2n} \log \delta^{-1} + \frac{2dL_{I,S}M_u}{\eta\tau_{pa}}
$$

where $\hat{\epsilon}_1 = 2m_{1}^{2} + 2\eta C + 2\lambda_{a}m_{1}^{2}/\tau_{pa} + 2\lambda_{a}^{2}dL_{I,S}M_u/\tau_{pa} + 2\lambda_{a}^{2}\log \mu_{y}^{2}$ and $\hat{\epsilon}_2 = \left(2 + 2k_{I}^{2}) \log \mu_{y}^{2} + \eta C + \frac{2k_{I}(1-\mu_{C})}{\tau_{pa}} + 2\lambda_{a}\lambda_{pa}dL_{I,S}M_u^2/\tau_{pa} + 2\lambda_{a}\lambda_{sa}\log \mu_{y}^{2}$ are constants dependent on $\{n, m, \mu_1, \mu_C, \mu_p, k'_I, \epsilon, L_{I,S}, L_{I}, M_u, d, C\}$. $C$ is a constant.

Theorem shows that our proposed method, with high probability $1 - \delta$, is with a bounded expected clustering risk on the unseen data. To summarize, the expected clustering risk of MCA is theoretically guaranteed in clustering tasks. Note that the margin $\mathcal{L}(g) - \mathcal{L}_{\tau}(g)$ is inversely proportional to $\mu_1$ and $\mu_C$ which reflect the neighborhood consistency in both image domain and cross-domain, and $\mu_p$ which reflects the prediction confidence, indicating that improving the neighborhood consistency in both image domain and cross-domain and prediction confidence reduces the expected risk of MCA. Meanwhile, the margin $\mathcal{L}(g) - \mathcal{L}_{\tau}(g)$
<table>
<thead>
<tr>
<th>Dataset</th>
<th>STL10</th>
<th>Cifar10</th>
<th>Cifar100-20</th>
<th>ImageNet-Dogs</th>
<th>Tiny-ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>19.2</td>
<td>12.5</td>
<td>6.1</td>
<td>13.0</td>
<td>10.5</td>
</tr>
<tr>
<td>SC</td>
<td>15.9</td>
<td>9.8</td>
<td>4.8</td>
<td>13.6</td>
<td>11.1</td>
</tr>
<tr>
<td>NMF</td>
<td>18.0</td>
<td>9.6</td>
<td>4.6</td>
<td>11.8</td>
<td>11.8</td>
</tr>
<tr>
<td>JULE</td>
<td>27.7</td>
<td>18.2</td>
<td>16.4</td>
<td>13.7</td>
<td>13.8</td>
</tr>
<tr>
<td>SAE</td>
<td>32.0</td>
<td>25.2</td>
<td>16.1</td>
<td>15.7</td>
<td>18.5</td>
</tr>
<tr>
<td>DAE</td>
<td>30.2</td>
<td>22.4</td>
<td>15.2</td>
<td>15.1</td>
<td>19.0</td>
</tr>
<tr>
<td>AE</td>
<td>30.3</td>
<td>25.0</td>
<td>16.1</td>
<td>16.5</td>
<td>18.5</td>
</tr>
<tr>
<td>VAE</td>
<td>28.2</td>
<td>20.0</td>
<td>14.6</td>
<td>15.2</td>
<td>17.9</td>
</tr>
<tr>
<td>DEC</td>
<td>35.9</td>
<td>27.6</td>
<td>18.6</td>
<td>18.5</td>
<td>19.5</td>
</tr>
<tr>
<td>ADC</td>
<td>53.0</td>
<td>-</td>
<td>-</td>
<td>16.0</td>
<td>-</td>
</tr>
<tr>
<td>DC</td>
<td>33.4</td>
<td>-</td>
<td>-</td>
<td>18.9</td>
<td>-</td>
</tr>
<tr>
<td>DAC</td>
<td>47.0</td>
<td>36.6</td>
<td>25.6</td>
<td>23.8</td>
<td>27.5</td>
</tr>
<tr>
<td>DDC</td>
<td>48.9</td>
<td>37.1</td>
<td>26.7</td>
<td>32.7</td>
<td>38.3</td>
</tr>
<tr>
<td>DCCM</td>
<td>48.2</td>
<td>37.6</td>
<td>26.2</td>
<td>25.7</td>
<td>35.2</td>
</tr>
<tr>
<td>IIC</td>
<td>59.6</td>
<td>49.6</td>
<td>39.7</td>
<td>33.7</td>
<td>52.6</td>
</tr>
<tr>
<td>PICA</td>
<td>71.3</td>
<td>61.1</td>
<td>53.1</td>
<td>49.6</td>
<td>42.9</td>
</tr>
<tr>
<td>GCC</td>
<td>78.8</td>
<td>68.4</td>
<td>63.1</td>
<td>56.7</td>
<td>42.9</td>
</tr>
<tr>
<td>CC</td>
<td>85.0</td>
<td>76.4</td>
<td>72.6</td>
<td>87.7</td>
<td>64.4</td>
</tr>
<tr>
<td>TCL</td>
<td>86.8</td>
<td>79.9</td>
<td>75.7</td>
<td>88.7</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 1: Clustering results on five benchmark datasets. The best results are highlighted in bold.

Comparisons with State-of-the-arts

Setup. We took the entire list of nouns in the WordNet dataset (Miller 1995) to form an initial semantic dataset for filtering which contains more than 82,000 nouns. To evaluate the effectiveness of our proposed method, we compare it with 22 state-of-the-art clustering methods on the five datasets, including k-means (MacQueen 1967), SC (Zelnik-Manor 2005), NMF (Cai et al. 2009), JULE (Yang, Parikh, and Batra 2016), SAE (Ng et al. 2011), DAE (Vincent et al. 2010), AE (Bengio et al. 2006), VAE (Kingma and Welling 2013), DEC (Xie, Girshick, and Farhadi 2016), ADC (Haeusser et al. 2018), DeepCluster (DC) (Caron et al. 2018), DAC (Chang et al. 2017a), DDC (Chang et al. 2019), DCCM (Wu et al. 2019), IIC (Ji, Vedaldi, and Henriques 2019), PICA (Huang, Gong, and Zhu 2020), GCC (Zhong et al. 2021), CC (Li et al. 2021), TCL (Li et al. 2022), SCN (Van Gansbeke et al. 2020), NNM (Dang et al. 2021) and SIC (Cai et al. 2023). We repeated the training five times independently on each dataset and reported their mean and standard deviation values.

Experiments and Analysis

In this section, experiments are conducted on five image benchmark datasets to validate the effectiveness of our proposed method.

Experimental Setup

Benchmarks and implementation details. We used the following five benchmark datasets in our experiment: STL10 (Coates, Ng, and Lee 2011), Cifar10 (Krizhevsky 2009), Cifar100-20 (Krizhevsky 2009), ImageNet-Dogs (Chang et al. 2017b) and Tiny-ImageNet (Le and Yang 2015).

Evaluation Metrics. We used three evaluation metrics to evaluate clustering results, including clustering Accuracy (ACC), Normalized Mutual Information (NMI) (McDaid, Greene, and Hurley 2011), and Adjusted Rand Index (ARI) (Hubert and Arabie 1985). For these metrics, a higher value means better performance.

is proportional to $k_r$ which reflects the neighborhood overlapping in the image domain, indicating that reducing the neighborhood imbalance (e.g., by setting a smaller number of neighbors $k_l$ or filtering neighborhoods to reduce neighborhood imbalance) also reduces the expected risk of MCA.
Table 2: Ablation studies of semantic space construction (UF: Uniqueness-based filtering, HF: Hierarchy-based filtering.)

<table>
<thead>
<tr>
<th>Loss Components</th>
<th>Result</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_L$</td>
<td>$L_{ia}$</td>
<td>$L_{pa}$</td>
<td>$L_{sa}$</td>
</tr>
<tr>
<td>ACC</td>
<td>NMI</td>
<td>ARI</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3: Ablation studies on ImageNet-Dogs.

Results. The comparison results with the state-of-the-art methods in terms of ACC, NMI, and ARI are presented in Table 1. From this table, we can observe that our method outperforms all other methods on five benchmark datasets. Especially, MCA improves ACC, NMI, and ARI by 2.8%, 1.3%, and 1.5% on Cifar100-20, and 6.6%, 3.3%, and 5.7% on ImageNet-Dogs, demonstrating that MCA better amend the incorrect alignments and thus achieve significant performance improvement. Especially for fine-grained images such as Imagenet-dogs whose categories are difficult to distinguish, our method can give useful clustering support information for images through the three alignment strategy, especially for fine-grained images like ImageNet-Dogs whose categories are difficult to distinguish.

Ablation Studies

Semantic space construction.

We first conduct an experiment on Cifar10 and ImageNet-Dogs to verify the effectiveness of our semantic space construction method and show the results in Table 2. The results demonstrate that hierarchy-based filtering can significantly enhance the semantic space compared to uniqueness-based filtering, especially when dealing with complex clusters that are challenging to distinguish, such as those in the ImageNet-Dogs dataset.

Figure 3: The average accuracy of 10 runs of pseudo-labels with epochs on Cifar100-20 and ImageNet-Dogs evolves.

Loss components effectiveness. We perform an ablation analysis on ImageNet-Dogs to measure the importance of four loss components in our model, i.e., image consistency loss $L_I$, instance-level alignment loss $L_{ia}$, prototype-level alignment loss $L_{pa}$, and semantic-level alignment loss $L_{sa}$. The results are shown in Table 3, indicating that each of the four components plays an important role. The integration of three cross-modal alignment strategies significantly enhances the clustering performance, even obtaining 28.1%, 23.8%, and 29.3% performance gains when all three strategies are simultaneously used. Among the three cross-modal alignment strategies, introducing semantic-level alignment yields the most significant improvement in clustering performance, indicating that operating at the semantic level can effectively address incorrect alignments. These results confirm the effectiveness of our proposed cross-modal alignment methods.

Comparison of three pseudo-label generation methods. In our method, semantic-level cross-modal alignment can be considered self-training with cross-modal pseudo-labels. To verify the effectiveness of our method, we compared three pseudo-label generation methods implemented in our framework: 1) Single Modal Pseudo-labeling (SMP) that directly generates one-hot pseudo-labels via the argmax operation on the soft cluster assignments only from images, 2) Prototype Mapping based Cross-modal Pseudo-labeling (PMCP), is the adjusted center-based method in (Cai et al. 2023)) that generates pseudo-labels from the prototype level alignments in the original CLIP, and 3) Ours that generates pseudo-labels by simultaneously learning relationships among images and neighboring texts, while also updating the image and text embeddings. The comparison results on Cifar100-20 and ImageNet-Dogs are shown in Figure 3, demonstrating that our method significantly outperforms the other two methods on both datasets. These results indicate that learning adaptive relationships at the semantic level substantially improves the quality of pseudo-labels.

Figure 4: Example of pseudo-label generation in MCA. The words below (on the right side of) the images are ground-truth/ neighboring labels and the red color indicates irrelevant texts. The blue block in the semantic probability indicates the class the left word is assigned to (with the largest probability).
Sensitivity Analysis

Figure 4 shows an example of our proposed pseudo-label generation method. We select an example from each of the three classes and three neighboring words for each image. Although these exist irrelevant neighboring words for an image, our method can identify irrelevant words and eliminate the affection of incorrect alignments in CLIP.

Sensitivity on neighborhood parameters $k_S$ and $k_p$ in cross-modal alignment. In our alignment method, $k_S$ controls the number of neighboring texts in the instance-level and semantic-level alignments, and $k_p$ controls the number of neighboring texts to recompute the semantic prototype in the prototype-level alignment. Figures 5a and 5b show that too large $k_S$ causes performance degeneration due to the introduction of irrelevant texts. Figures 5c and 5d show that $k_p$ does not change the performance too much.

Sensitivity on trade-off parameters $\eta$, $\lambda_a$, $\lambda_{pa}$ and $\lambda_{sa}$. Figure 6 shows the sensitivity analysis of trade-off parameters $\eta$, $\lambda_a$, $\lambda_{pa}$ and $\lambda_{sa}$. We can see that the performance of our method improves with increasing values of $\eta$, $\lambda_{pa}$, and $\lambda_{sa}$. Notably, our method appears to be more sensitive to changes in $\lambda_{sa}$ than $\lambda_{pa}$ on ImageNet-Dogs.

Compare with Zero-shot Learning

We compared MCA with two zero-shot learning methods, i.e., CLIP and MUST. Compared to CLIP, MUST improve CLIP by 3.3% on Caltech101 and 18% on UCF101. Compared to MUST, our method deteriorates MUST by -8.2% on Caltech101 and -8.5% on UCF101. Although the results show that our method performs worse than CLIP and MUST, our method has the advantage of not requiring class names as input, which is necessary for zero-shot learning. This makes our approach more versatile and applicable to a broader range of real-world scenarios.

Conclusion

We have proposed a novel method to address the incorrect alignments in CLIP for image clustering. Our method includes the construction of a proper semantic space and a multi-level cross-modal alignment approach for aligning images and texts in downstream tasks at three levels. Theoretical results have shown interesting insights, and experimental results have demonstrated the superiority of our method. However, we acknowledge that our method may not be as cost-effective as SIC, as it involves three types of alignments. The proper setting of hierarchy levels also remains a challenge that needs further investigation. Additionally, our method exhibits lower performance compared to MUST, primarily due to the absence of class names. For future work, we will focus on enhancing the performance of our method and exploring new avenues for improvement. Leveraging the theoretical results to augment our method holds promise and will be a key research direction we pursue.
Acknowledgments
This work is jointly supported by Major Project of the New Generation of Artificial Intelligence under Grant no.2018AAA0102900; in part by NSFC under Grant no. 92270122 and no.62206179; and in part by Guangdong Provincial Natural Science Foundation under Grant no. 2023A1515012584 and no.2022A1515010129; and in part by the Shenzhen Research Foundation for Basic Research, China, under Grant JCYJ20210324093000002; and in part by University Stability Support program of Shenzhen under Grant no.20220811121315001.

References


Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, 4904–4916. PMLR.


