TiMix: Text-Aware Image Mixing for Effective Vision-Language Pre-training

Chaoya Jiang¹, Wei Ye¹*, Haiyang Xu²*, Qinghao Ye², Ming Yan², Ji Zhang², Shikun Zhang¹

¹National Engineering Research Center for Software Engineering, Peking University, Beijing, China
²Alibaba Group, Hangzhou, China
wye@pku.edu.cn, shuofeng.xhy@alibaba-inc.com

Abstract

Self-supervised Multi-modal Contrastive Learning (SMCL) remarkably advances modern Vision-Language Pre-training (VLP) models by aligning visual and linguistic modalities. Due to noises in web-harvested text-image pairs, however, scaling up training data volume in SMCL presents considerable obstacles in terms of computational cost and data inefficiency. To improve data efficiency in VLP, we propose Text-aware Image Mixing (TiMix), which integrates mix-based data augmentation techniques into SMCL, yielding significant performance improvements without significantly increasing computational overhead. We provide a theoretical analysis of TiMix from a mutual information (MI) perspective, showing that mixed data samples for cross-modal contrastive learning implicitly serve as a regularizer for the contrastive loss. The experimental results demonstrate that TiMix exhibits a comparable performance on downstream tasks, even with a reduced amount of training data and shorter training time, when benchmarked against existing methods. This work empirically and theoretically demonstrates the potential of data mixing for data-efficient and computationally viable VLP, benefiting broader VLP model adoption in practical scenarios. Our code is available on https://github.com/chaoyajiang/TiMiX/tree/main.

Introduction

Vision-Language Pre-training (VLP) exploits large-scale image-text pairs without annotations via self-supervised learning (Syed, Gaol, and Matsu 2021; Liu et al. 2020), achieving tremendous success on a wide range of cross-modal downstream tasks (Chen et al. 2020; Huang et al. 2020; Li et al. 2020; Yu et al. 2021; Li et al. 2021; Zhang et al. 2021; Kim, Son, and Kim 2021; Li et al. 2022a; Xu et al. 2021; ). More recently, Self-supervised Multi-modal Contrastive Learning (SMCL) has emerged as a significant advancement in the VLP community (Li et al. 2022b; Radford et al. 2021b; Li et al. 2021, 2022a; Zeng, Zhang, and Li 2021), facilitating the learning of cross-modal representations from image-text pairs by aligning visual and linguistic modalities.

Recent studies (Li et al. 2021, 2022b; Jiang et al. 2023c) have found that SMCL-based models pre-trained on web-harvested data often suffer from data inefficiency since image captions frequently contain words that are unrelated to the image content or only capture partial information. One common strategy is increasing the scale of the training to alleviate the negative impacts of noisy data samples (Radford et al. 2021b; Jia et al. 2021). A typical example is CLIP (Radford et al. 2021b), which utilizes a massive dataset of 400 million image-text pairs obtained through web crawling. Though it demonstrated promising results in enhancing the cross-modal capabilities of models, scaling up datasets presents a challenge due to the high computational cost. For example, CLIP requires an estimated 3584 GPU (V100) days for pertaining, a demand that is financially prohibitive under a constrained budget. Other researchers exploit soft labels (Li et al. 2021) or regenerate image captions (Li et al. 2022b) to mitigate the impact of noisy data, yet with unsatisfactory performance improvement or substantial additional computation.

In this paper, we present a novel perspective of data mixing to tackle data inefficiency in VLP. We hypothesize that an image could exhibit multiple distinct views, each potentially associated with a different textual caption. These diverse textual descriptions align with specific views that capture various aspects of the image’s semantic information. Building upon this hypothesis, we introduce Text-aware Image Mixing
we design a patch-text alignment (PTA) pre-training task, specifically adopting the CutMix (Yun et al. 2019) approach. Given a pre-trained object detector and then train the cross-modal representation and enhance data efficiency.

We theoretically analyze TiMix from a mutual information (MI) maximization perspective and find that mixed data samples implicitly provide a regularizer for the contrastive learning loss function. This regularizer keeps the model from overfitting to partially aligned image-text pairs during the contrastive learning process, thereby mitigating the negative impact of noisy data. Empirically, by incorporating TiMix into existing VLP models, we can observe consistent performance improvement on common vision-language downstream tasks, including Visual Question Answering (VQA), Cross-modal Retrieval, Natural Language for Visual Reasoning (NLVR) and Image Captioning, with small additional computational cost during training.

In summary, our contributions are:

• We take the first step to introduce mix-based data samples into vision-language pre-training. With a novel patch-text alignment pre-training task, mixed images are created in a CutMix style based on the matching degree of their patches and captions, serving as high-quality data for cross-modal contrastive learning.

• We theoretically prove that mixed data samples implicitly provide a regularizer for cross-modal contrastive learning, facilitating mutual information optimization for potentially partially-aligned image-text pairs.

• Experimental findings illustrate that TiMix delivers robust performance, significantly enhancing data efficiency while maintaining cost-effectiveness during the pre-training phase. For example, as shown in Figure 1, TiMix achieves comparable downstream task performance by training on 40% of the data in 43.8% of the training time, compared to a recent robust VLP model mPLUG.

Related Work
Vision-Language pre-training
Recent years have seen significant success for large-scale pre-trained vision-language models (Tan and Bansal 2019; Chen et al. 2020; Huang et al. 2020; Li et al. 2020; Yu et al. 2021; Li et al. 2021; Wang et al. 2021b; Li et al. 2022a; Zhang et al. 2021; Jiang et al. 2023a,b) in a variety of cross-modal tasks. Current approaches to VLP can be broadly divided into two categories in terms of visual representation extraction. The first category is detector-based VLP methods (Li et al. 2019; Tan and Bansal 2019; Li et al. 2020; Chen et al. 2020; Yu et al. 2021; Fang et al. 2021). These methods primarily adopt a two-step training pipeline: they extract visual features using a pre-trained object detector and then train the cross-modal pre-training model to align text and visual features. The main challenge for these methods is to balance effectiveness and efficiency. The second category consists of more recent CNN-based (Xu et al. 2021) or ViTs-based (Li et al. 2021; Kim, Son, and Kim 2021; Radford et al. 2021a) methods, especially patch-based ViT. These methods eliminate the need for a complex object detector in feature extraction, enabling end-to-end VL learning. Furthermore, Self-supervised Multi-modal Contrastive Learning (SMCL) has lately sparked significant advancements (Li et al. 2022b; Radford et al. 2021b; Yao et al. 2021; Li et al. 2021, 2022a) by conducting cross-modal alignment. SMCL consists of image-to-text and text-to-image contrastive learning, e.g., with the InfoNCE (Oord, Li, and Vinyals 2018) loss.

Mixed Data Augmentation
Mixup (Zhang et al. 2017) is a widely used data augmentation technique in Computer Vision, which involves training by convexly combining the input image and its corresponding label. CutMix (Yun et al. 2019), on the other hand, is a specific case of Mixup and can be seen as a pixel-level Mixup method that utilizes binary masks. Recent developments (Uddin et al. 2020; Liu et al. 2022; Walawalkar et al. 2020) in Mixup methods have focused on effectively leveraging saliency information and performing mixing at the image feature level. The motivation behind saliency-based Mixup methods is to preserve salient regions when blending images, ensuring that sufficient information is retained and more natural feature representations are learned. SaliencyMix (Uddin et al. 2020) employs various saliency detectors to directly extract salient regions from the images. Co-Mixup (Kim et al. 2021) aims to maximize the gradient-based saliency while encouraging diversity in the mixed images' hypermodules. SuperMix (Dabouei et al. 2020) utilizes supervised signals to mix input images based on saliency. However, the majority of existing Mixup methods have mainly been applied in the CV. In contrast, we introduce Mixup techniques to the multimodal domain for the first time, resulting in notable advancements.

Method
Our approach exhibits similarities to the Cutmix-style method. Specifically, our method involves pairing two images within a batch, denoted as $I^x$ and $I^y$, along with their corresponding texts, $T^x$ and $T^y$. Assuming $I^y$ as the source image, we extract the region $R^y$ in the source image $I^y$ that exhibits the highest text-relevant score with $T^y$. Subsequently, we replace the corresponding region $R^y$ in the target image $I^x$, which possesses the lowest text-relevant score with $T^x$, with the extracted region $R^y$. This process yields a mixed image denoted as $I^x:y$. To determine the soft labels for the mixed image and text, we consider the proportion of the mixed and cropped regions. Additional details regarding this process are provided in the following section.

Text-aware Region Mixing
As depicted in Figure 2, considering two pairs of image-text denoted as $I^x, T^x$ and $I^y, T^y$ respectively, let us designate $I^x$ as the target image and $I^y$ as the source image. Initially, suppose the image size is $H \times W$, the target image $I^x$ and the source image $I^y$ are divided into distinct, non-overlapping patches of size $P \times P$, resulting in a total of $\frac{H^2}{P^2} \times \frac{W^2}{P^2}$ patches. Subsequently, these patches are fed into
The key component of TiMix is the text-aware patch predictor which needs to predict text-relevant scores between the image patches and input text. As shown in Figure 2, the patch predictor is a Multi-Layer Perceptron (MLP) that contains three linear layers and is used to predict the alignment score between patches and the input text.

As the lack of fine-grained patch-text labels to train the text-aware patch predictor, in this sub-section, we propose to convert object-level signals into patch-level ones and introduce a novel pre-training task named Patch Text Alignment which facilitates the patch predictor training and drives our model to learn the fine-grained patch-text alignment. For object objection and visual grounding datasets like COCO(Lin et al. 2014) and VG(Krishna et al. 2016), the object and region generally be paired with a class label or text description. Therefore, we can transfer every object class label to a text description based on a template such as “This is a [class label]”. Thus, for each (object/region) bounding box in an image, we can construct a text description for it. Then, we transform the bounding box annotations to the patch-level.

The soft label of the mixed image \( \hat{I}^{x,y} \) to text \( t^y \) is calculated as follow:

\[
s^y = \frac{hwP^2}{HW}
\]  (5)

Figure 2: The subfigure (a) illustrates the process of TiMix, where two image-text pairs are utilized. Subfigure (b) depicts the architecture of the Text-aware patch predictor.
Contrastive Learning Based on TiMix

We introduce how to apply TiMix to the mixed image \( \hat{y}^{x:y} \). By repeating this process, we acquire \( N \) mixed images. As shown in Figure 3, we select a mixed image sample \( \hat{y}^{x:y} \) as the anchor. The two texts \( T^x \) and \( T^y \) associated with the source image and target image are regarded as positive samples. Following the aforementioned procedure, we compute the soft labels of the anchor image to these two positive samples. The remaining texts are considered negative samples. Suppose the extracted global vision representation of the mixed image is denoted as \( \hat{v}^{x:y} \) and there are two positive text samples, while the other mixed samples in the batch are regarded as negative samples. The label assigned to the text anchor with respect to these two positive image samples is the same as the labels used in image-to-text contrastive learning.

A Mutual Information Maximization Perspective

In this section, we provide evidence and explanations for our method from the perspective of maximizing mutual information. Following the definition in (Oord, Li, and Vinyals 2018) in the context of image-to-text contrastive learning, the similarity function \( f(v_i, t_k) \) in Equation 8 can be utilized to model the density ratio, which preserves the mutual information between the image \( v_i \) and the text \( t_k \). We rewrite the \( f(v_i, t_k) \) to \( \frac{P(t_k|v_i)}{P(t_k)} \). Then, given a batch of unmixed image-text pairs, the vanilla contrastive learning loss \( \mathcal{L}^v \) satisfies the following inequality:

\[
\mathcal{L}^v = - \mathbb{E} \log \left[ \frac{P(t_k|v_i)}{P(t_k)} + \sum_{k \neq i} \frac{P(t_k|v_i)}{P(t_k)} \right] 
\]

where \( I(t_i, v_i) \) denotes the mutual information between \( t_i \) and \( v_i \). The detailed proof can be found in Appendix A. Based on inequality 9, we can get the lower bound of \( I(t_i, v_i) \) as:

\[
I(t_i, v_i) \geq \log(N) - \mathcal{L}^v 
\]

With a similar derivation, we can get another inequality about the image-to-text contrastive learning loss in TiMix as fol-
We evaluate TiMix with two well-known VLP models ALBEF and mPLUG (denoted as ALBEF-TiMix and mPLUG-TiMix) on four vision-language downstream tasks: visual question answering (VQA2.0 (Agrawal et al., 2015)), natural language for visual reasoning (NLVR2 (Suhr et al., 2018)), image-text retrieval (Flickr30K (Plummer et al., 2015), COCO (Lin et al., 2014)), and language for visual reasoning (NLVR2 (Suhr et al., 2018)).

Table 1: Evaluation results on VQA2.0 and NLVR2. More details about comparison models are in Appendix E.

The details of this derivation can be found in Appendix A. Combining inequality 10 and 13 provides us with an improvement of data efficiency.

**Experiment**

Following the previous works (Li et al. 2021) and (Li et al. 2022a), we use the same pre-training dataset with 14M images with texts, which includes two in-domain datasets (MS COCO (Lin et al. 2014) and Visual Genome (Krishna et al., 2016), and three web out-domain datasets (Conceptual Captions (Sharma et al. 2018a), Conceptual 12M (Changpinyo et al., 2021a), SBU Captions (Ordonez, Kulkarni, and Berg 2011)). Please refer to Appendix C to see more detail about the pre-training dataset and pre-training setting.

**Overall Performance**

We evaluate TiMix with two well-known VLP models ALBEF and mPLUG (denoted as ALBEF-TiMix and mPLUG-TiMix) on four vision-language downstream tasks: visual question answering (VQA2.0 (Agrawal et al. 2015)), natural language for visual reasoning (NLVR2 (Suhr et al. 2018)), image-text retrieval (Flickr30K (Plummer et al. 2015), COCO (Lin et al. 2014)), and language for visual reasoning (NLVR2 (Suhr et al. 2018)). Our baselines cover 16 VLP models, detailed in Appendix E. In our experiments, we only re-implement the base version of ALBEF (Li et al. 2021) and mPLUG (Li et al. 2022a). We will first analyze their overall performances on these tasks. The fine-tuning hyper-parameters and the details of downstream tasks are described in Appendix D.

**Visual Question Answering**

The VQA task requires the model to answer natural language questions given an image. Following the approach of (Li et al. 2021), we treat VQA as an answer generation problem. We evaluated our models by submitting the results to the evaluation server and report the test-dev and test-std scores in Table 1. The VLP models equipped with TiMix demonstrate improved performance on the VQA task compared to the models without TiMix. These results highlight the significant improvements achieved by TiMix. Additionally, our mPLUG-TiMix model trained on 14M data outperforms other baseline models which provides further evidence of the effectiveness of our method.

**Natural Language for Visual Reasoning**

The NLVR2 (Suhr et al. 2018) task requires the model to predict whether a sentence accurately describes a pair of images, which is a binary classification task. For ALBEF-TiMix and mPLUG-TiMix, we follow (Li et al. 2021) and use two cross-attention layers to process the two input images: their outputs are merged and fed into a Feed Forward Network (FFN). As shown in Table 1, pre-trained with 14M, mPLUG-TiMix can obtain competitive performances to the SOTA models.

---

https://eval.ai/web/challenges/challenge-page/830/leaderboard
Table 2: Evaluation results of image-text retrieval on Flickr30K (Plummer et al. 2015) and COCO datasets (Lin et al. 2014).

<table>
<thead>
<tr>
<th>Models</th>
<th># Pre-train data</th>
<th>MSCOCO (5K test set)</th>
<th>Flickr30K (1K test set)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TR</td>
<td>IR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R@1 R@5 R@10 R@1 R@5 R@10</td>
<td>R@1 R@5 R@10 R@1 R@5 R@10</td>
</tr>
<tr>
<td>E2E-VLP</td>
<td>4M</td>
<td>95.9 97.9 65.1 86.3 91.8</td>
<td>- - - - - -</td>
</tr>
<tr>
<td>OSCAR</td>
<td>4M</td>
<td>85.7 96.8 98.6 94.9 99.8 100.0</td>
<td>3.3M 96.7 98.0 65.2 97.4 99.9</td>
</tr>
<tr>
<td>ViViT</td>
<td>5.6M</td>
<td>- - - - - -</td>
<td>- - - - - -</td>
</tr>
<tr>
<td>ViL-Bert</td>
<td>3.3M</td>
<td>- - - - - -</td>
<td>- - - - - -</td>
</tr>
<tr>
<td>UNITER</td>
<td>4M</td>
<td>77.6 94.3 97.2 60.7 84.3 90.5</td>
<td>95.9 99.8 100.0 85.6 97.5 99.8</td>
</tr>
<tr>
<td>METER</td>
<td>4M</td>
<td>78.8 95.2 97.6 61.3 85.2 91.0</td>
<td>96.7 99.8 100.0 86.4 97.2 99.0</td>
</tr>
<tr>
<td>VLMo</td>
<td>4M</td>
<td>80.5 95.3 97.2 63.3 85.4 91.5</td>
<td>96.8 99.8 100.0 86.2 97.6 99.8</td>
</tr>
<tr>
<td>ViLT</td>
<td>4M</td>
<td>80.6 94.8 97.1 63.9 85.5 91.2</td>
<td>96.5 99.8 100.0 86.3 97.2 98.9</td>
</tr>
<tr>
<td>ALBEF</td>
<td>4M</td>
<td>73.1 91.4 96.0 56.8 81.5 89.2</td>
<td>94.3 99.4 99.8 82.8 96.7 98.4</td>
</tr>
<tr>
<td>ALBEF-TiMix</td>
<td>4M</td>
<td>76.4 93.7 96.6 60.4 83.2 90.3</td>
<td>95.1 99.8 100.0 84.2 97.3 98.6</td>
</tr>
<tr>
<td>ALBEF</td>
<td>14M</td>
<td>77.6 94.3 97.2 60.7 84.3 90.5</td>
<td>95.9 99.8 100.0 85.6 97.5 99.8</td>
</tr>
<tr>
<td>ALBEF-TiMix</td>
<td>14M</td>
<td>78.8 95.2 97.6 61.3 85.2 91.0</td>
<td>96.7 99.8 100.0 86.4 97.2 99.0</td>
</tr>
<tr>
<td>mPLUG</td>
<td>4M</td>
<td>80.5 95.3 97.2 63.3 85.4 91.5</td>
<td>96.8 99.8 100.0 86.2 97.6 99.8</td>
</tr>
<tr>
<td>mPLUG-TiMix</td>
<td>4M</td>
<td>80.6 94.8 97.1 63.9 85.5 91.2</td>
<td>96.5 99.8 100.0 86.3 97.2 98.9</td>
</tr>
<tr>
<td>mPLUG</td>
<td>14M</td>
<td>80.6 94.8 97.1 63.9 85.5 91.2</td>
<td>96.5 99.8 100.0 86.3 97.2 98.9</td>
</tr>
<tr>
<td>mPLUG-TiMix</td>
<td>14M</td>
<td>82.3 95.8 98.0 65.2 87.0 92.1</td>
<td>97.2 99.8 100.0 87.9 97.8 99.0</td>
</tr>
</tbody>
</table>

Figure 4: The visualization of VQA accuracy and Pre-training time per epoch

(a) VQA accuracy. (b) Pre-training time per epoch

Image-Text Retrieval We conduct experiments for both image-to-text retrieval (TR) and text-to-image retrieval (IR) on MSCOCO (Lin et al. 2014) and Flickr30K (Plummer et al. 2015) datasets. As shown in Table 2, pre-trained with 14M images, mPLUG-TiMix outperforms all existing methods on both datasets which even achieves better performance than BLIP with 129M. In addition, all models equipped with TiMix show significant improvements compared to their counterparts without TiMix.

Image Captioning Following (Li et al. 2022a), we fine-tune mPLUG/ mPLUG-TiMix with cross-entropy loss and then with CIDEr optimization for an additional 5 epochs. Our experiments, as shown in Table 3, unequivocally illustrate the superiority of mPLUG-TiMix over mPLUG alone. Notably, mPLUG-TiMix achieves performance levels that are comparable to those of SOTA models.

Ablation Study

We conducted ablation studies to examine the impact of the Pre-training task Patch Text Alignment (PTA) and mix-based data augmentation. Specifically, we investigated the effects of removing the PTA task while keeping the mix-based data augmentation for contrastive learning (w/o PTA). Without the PTA task, the Text-aware Patch Predictor cannot be optimized effectively, so we replaced it with a simple strategy where we follow the way of CutMix (Yun et al. 2019) and randomly sample the region in the image. In Table 4, we can observe that without the text guidance (w/o PTA), randomly mixing the image regions leads to a negligible improvement in accuracy on VQA and NLVR compared to the baseline model mPLUG (w/o TiMix). This demonstrates the effectiveness of our TiMix approach in leveraging text guidance for improved performance. In the case denoted as w/o Mix, we exclude the mix-based data augmentation method and only retain the PTA task. As presented in Table 4, we observed that utilizing only the PTA task still leads to a notable improvement in performance. This finding suggests that PTA enables our model to learn fine-grained cross-modal semantic alignment, thereby enhancing performance, although the improvement may not be substantial.

Impact of Pre-training Data

To gain a deeper comprehension of the influence of pre-training data size on the efficacy of TiMix , we conducted pre-training with data sizes of 4M, 6M, 8M, 10M, and 14M. These datasets were further augmented using two different mixing strategies: TiMix , Mixup (Zhang et al. 2017) and CutMix (Yun et al. 2019). Figure 4 (a) showcases the VQA results for various mix strategies, as well as the baseline model mPLUG, which does not employ mix-based augmentation.
Table 3: Evaluation Results on image captioning on COCO Karpathy test split (Karpathy and Fei-Fei 2015). B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-train Data</th>
<th>Cross-entropy Optimization</th>
<th>COCO Caption</th>
<th>CIDEr Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B@4</td>
<td>M</td>
<td>C</td>
</tr>
<tr>
<td>E2E-VLP</td>
<td>4M</td>
<td>36.2</td>
<td>117.3</td>
<td>-</td>
</tr>
<tr>
<td>OSCAR</td>
<td>6.5M</td>
<td>36.5</td>
<td>133.7</td>
<td>23.1</td>
</tr>
<tr>
<td>VinVL</td>
<td>5.65M</td>
<td>38.5</td>
<td>130.8</td>
<td>23.4</td>
</tr>
<tr>
<td>BLIP</td>
<td>14M</td>
<td>38.6</td>
<td>129.7</td>
<td>-</td>
</tr>
<tr>
<td>LEMON</td>
<td>200M</td>
<td>40.6</td>
<td>135.7</td>
<td>23.5</td>
</tr>
<tr>
<td>UFO</td>
<td>4M</td>
<td>38.7</td>
<td>131.2</td>
<td>23.3</td>
</tr>
<tr>
<td>SimVLM</td>
<td>1.8B</td>
<td>39.0</td>
<td>134.8</td>
<td>24.0</td>
</tr>
<tr>
<td>mPLUG</td>
<td>4M</td>
<td>39.5</td>
<td>132.6</td>
<td>23.2</td>
</tr>
<tr>
<td>mPLUG-TiMix</td>
<td>4M</td>
<td>40.7</td>
<td>134.8</td>
<td>23.9</td>
</tr>
<tr>
<td>mPLUG</td>
<td>14M</td>
<td>41.4</td>
<td>136.8</td>
<td>23.6</td>
</tr>
<tr>
<td>mPLUG-TiMix</td>
<td>14M</td>
<td>42.2</td>
<td>138.3</td>
<td>24.1</td>
</tr>
</tbody>
</table>

Table 4: The results of the ablation studies where we report the VQA and NLVR performance of various model variants. PT refers to Pretraining Time

Notably, TiMix consistently exhibits superior performance across the entire range. This observation suggests that TiMix not only enhances data efficiency in scenarios with limited data but also delivers substantial performance gains as the dataset size expands. From Figure 4 (a), we have observed that CutMix and Mixup provides only limited improvements in model performance. This indicates that the conventional Mixup approach does not significantly enhance the model’s performance. Furthermore, it demonstrates the effectiveness of our approach.

Data Efficiency of TiMix

To explore the effects of TiMix on the additional computational costs during pre-training, we conducted experiments to measure the training time per epoch for mPLUG trained without any data augmentation, as well as mPLUG utilizing TiMix, CutMix (Yun et al. 2019) and Mixup (Yun et al. 2019) for contrastive learning with additional data augmentation. As shown in Figure 4 (b), we recorded the corresponding training times for various data scales. We found that although TiMix, CutMix and mixup introduce some additional training time and computational overhead, the increase in overhead is not significant. Compared to the baseline model mPLUG, mPLUG-TiMix achieves significant improvements in model performance with relatively less computational cost. For example, to achieve the same performance as mPLUG-TiMix on 4M data size, the baseline mPLUG would require scaling the pre-training data to 10M, resulting in much higher computational overhead. This demonstrates the data efficiency of TiMix.

Conclusion

This paper addresses the challenges of scaling up training data volume in Self-supervised Multi-modal Contrastive Learning (SMCL) for Vision-Language Pre-training (VLP) models. We have introduced Text-aware Image Mixing (TiMix) as a solution to improve data efficiency in VLP by integrating mix-based data augmentation techniques into SMCL. Through a theoretical analysis from a mutual information (MI) perspective, we have theoretically shown that well-mixed data samples serve as a regularizer for the classical InfoNCE loss, empirically resulting in significant performance improvements without incurring excessive computational overhead and thereby significantly improving data efficiency in VLP.

Acknowledgments

This research is supported by the National Key Research And Development Program of China (No. 2021YFC3340101).

References


Bi, B.; Li, C.; Wu, C.; Yan, M.; Wang, W.; Huang, S.; Huang, F.; and Si, L. 2020. Palm: Pre-training an autoencod-


Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In ICLR.


Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In NIPS.


