NLIP: Noise-Robust Language-Image Pre-training

Runhui Huang¹, Yanxin Long¹, Jianhua Han², Hang Xu², Xiwen Liang¹, Chunjing Xu², Xiaodan Liang¹⁺

¹ Shenzhen campus of Sun Yat-sen University,
² Huawei Noah’s Ark Lab

{huangrh9, longyx9}@mail2.sysu.edu.cn, hanjianhua4@huawei.com, chromexbjxh@gmail.com,
liangcici5@gmail.com, xuchunjing@huawei.com, xdliang328@gmail.com

Abstract

Large-scale cross-modal pre-training paradigms have recently shown ubiquitous success on a wide range of downstream tasks, e.g., zero-shot classification, retrieval and image captioning. However, their success highly relies on the scale and quality of web-crawled data that naturally contain much incomplete and noisy information (e.g., wrong or irrelevant content). Existing works either design manual rules to clean data or generate pseudo-targets as auxiliary signals for reducing noise impact, which do not explicitly tackle both the incorrect and incomplete challenges at the same time. In this paper, to automatically mitigate the impact of noise by solely mining over existing data, we propose a principled Noise-robust Language-Image Pre-training framework (NLIP) to stabilize pre-training via two schemes: noise-harmonization and noise-completion. First, in noise-harmonization scheme, NLIP estimates the noise probability of each image-text pair according to the memorization effect of cross-modal transformers, then adopts noise-adaptive regularization to harmonize the cross-modal alignments with varying degrees. Second, in noise-completion scheme, to enrich the missing object information of text, NLIP injects a concept-conditioned cross-modal decoder to obtain semantic-consistent synthetic captions to complete noisy ones, which uses the retrieved visual concepts (i.e., objects’ names) for the corresponding image to guide captioning generation. By collaboratively optimizing noise-harmonization and noise-completion schemes, our NLIP can alleviate the common noise effects during image-text pre-training in a more efficient way. Extensive experiments show the significant performance improvements of our NLIP using only 26M data over existing pre-trained models (e.g., CLIP, BLIP) on 12 zero-shot classification datasets (e.g., +8.6% over CLIP on average accuracy), MSCOCO image captioning (e.g., +1.9 over BLIP trained with 129M data on CIDEr) and zero-shot image-text retrieval tasks.

Introduction

Vision-Language Models (VLMs) (Yao et al. 2021; Radford et al. 2021; Li et al. 2021; Jia et al. 2021; Li et al. 2022a) pre-trained with image-text pairs has shown its extraordinary zero-shot transfer abilities in different downstream tasks, including zero-shot classification (Radford et al. 2021; Yao et al. 2021), image-text retrieval (Radford et al. 2021; Yao et al. 2021), image captioning (Wang et al. 2021) and text-to-image generation (Patashnik et al. 2021), etc. Previous works (Radford et al. 2021; Li et al. 2022a) show that the downstream performance of VLMs highly relies on the scale or the quality of pre-training image-caption pairs. However, considering the prohibitive expense of acquiring high-quality annotated image-caption datasets (Lin et al. 2014), current paradigms resort to collecting increasingly larger sizes of unlabeled image-text datasets (Thomee et al. 2016; Sharma et al. 2018), largely overlooking the prevalent noise in the web. They thus lead to the heavier computation burden and make the pre-training process severely unstable due to the negative impact of noise.

To leverage the advantages of both quality and scale, several attempts have been made to mitigate the negative impact of noisy pairs. On the one hand, some filtering and post-processing procedures (Sharma et al. 2018; Changpinyo et al. 2021; Jia et al. 2021) have been designed to clean up the large-scale unlabeled data for pre-training. On the other hand, few works explore automatic ways during training. For example, ALBEF (Li et al. 2021) resorts to a momentum model to generate pseudo-targets as additional supervision. BLIP (Li et al. 2022a) uses a filter to remove the noisy data rectified by the similarity of image-text pairs and a captioner to regenerate texts. NCR (Huang et al. 2021) utilizes the loss...
distribution to divide clean samples and noisy samples and then rectify the labels by model predictions. However, unlabeled “noise” data often naturally appear with either incorrect text descriptions or incomplete ones (e.g., missing descriptions of some object concepts), where none of the existing works consider automatically alleviating both of them within one framework. Here, we aim to achieve noise-robust learning from two aspects: self-diagnosing incorrect vs. correct pairs and harmonizing the loss; self-generating and selecting confident captions with enriched concepts.

To fully utilize the entire image-caption pairs including the noisy ones, we introduce a principled Noise-robust Language-Image Pre-training framework (NLIP) to stabilize pre-training by noise-harmonization and noise-completion schemes: (a) Noise-harmonization, where NLIP learns to harmonize the cross-modal alignment and adopts noise-adaptive regularization for each pair based on the estimated noisy probability. Specifically, Arpit et al. (2017) suggests that deep network tends to fit the easy (i.e., clean) samples first and then the noisy ones. Based on the memorization effect of cross-modal transformers, NLIP first estimates the noise probability for each pair, then applies a noise-adaptive regularization on the image-text contrastive loss to avoid over-fitting to the noisy data (shown in Fig.1(a)). This scheme pulls the embeddings of the image and caption in the clean pair more tightly than the one with a higher noisy probability. (b) Noise-completion, where NLIP employs a concept-conditioned cross-modal decoder to synthesize semantic-consistent captions to replace the detrimental noisy texts. Specifically, to guide the caption generation procedure via providing prior information about the existing objects, we first retrieve the visual concepts (i.e., names of existing objects) for each image via a pre-trained VLM. Then these visual concepts and the image are fed into an additional caption head to generate the enriched descriptions for each noisy pair to substitute the noisy caption (shown in Fig.1(b)). Furthermore, inspired by He et al. (2021), we explore enhancing the visual encoder via randomly masking the input image tokens and then reconstructing them, which can help reduce the computation cost during training and boost visual embedding by maintaining low-level visual information.

Experimental results show that NLIP achieves significant performance on several downstream tasks, including zero-shot classification, zero-shot image-to-text/text-to-image retrieval and image-captioning tasks. Our NLIP outperforms CLIP (Radford et al. 2021) by 8.6% in terms of average accuracy on 12 zero-shot classification datasets. With respect to image captioning, NLIP is superior to existing image captioning methods that are trained with substantially more data, e.g., 1.9 over BLIP (Li et al. 2022a) trained with 129M image-text pairs in terms of CIDEr on MSCOCO. For zero-shot image-text retrieval tasks, NLIP surpasses CLIP by 28.7% in terms of R@1 on Flickr30k.

Related Work

Vision Language Pre-training (VLP) models recently garner increasing attention as the surprisingly superior performances on diverse zero-shot downstream tasks. They propose to learn semantic alignments across image and language modalities by pre-training on large-scale data which brings strong performance benefits in downstream tasks (e.g., zero-shot classification, zero-shot retrieval, image caption). Existing VLP models often appear with either encoder-only or encoder-decoder architectures. The encoder-only architectures (Radford et al. 2021; Jia et al. 2021; Yao et al. 2021; Yuan et al. 2021; Mu et al. 2021; Li et al. 2022b; You et al. 2022) aim to align the visual features with textual features in a common cross-modal semantic space. The encoder-decoder architectures (Wang et al. 2021; Li et al. 2022a) employ autoregressive Language Modeling (LM) (e.g., image captioning, text-grounded image generation) to supervise the decoder and excel in generation-related downstream tasks. Despite the nature merits in data diversity, the large-scale web-crawled image-text pairs contain much noise (i.e., incomplete or even error information) (Thomee et al. 2016; Changpinyo et al. 2021). Some works attempt to mitigate the impact in two aspects. From the data perspective, some strict rules are used to clean up the data (Sharma et al. 2018; Changpinyo et al. 2021; Jia et al. 2021). From the modeling perspective, ALBEF (Li et al. 2021) adopts momentum models to generate pseudo-targets as additional supervision; BLIP (Li et al. 2022a) presents a filter to remove the noisy data rectified by the similarity of image-text pairs and a captioner to regenerate the corresponding web texts. However, they have not explicitly stabilized and harmonized the pre-training objectives by reevaluating noisy data in a soft way. In this work, we alleviate the noisy impact by simultaneously addressing incorrect and incomplete image-text pairs. Two novel noise-harmonization and noise-completion schemes are collaborative to achieve noise-robust pre-training.

Noisy Data Learning has been a long-standing research area to cope with the noise in training data, practically all of which are applied to the classification task. Existing studies (Song et al. 2020) frequently use robust architecture design, regularization, loss modification, or sample selection strategies to limit the detrimental impact of noisy labels. Here we discuss the last three techniques, which are the most relevant to our model. First, the regularization enforces the networks to over-fit less to false-labeled examples explicitly or implicitly, e.g., label smoothing (Pereyra et al. 2017; Lukasik et al. 2020) avoids over-fitting by preventing the networks from assigning full probabilities to noisy data samples. Second, the loss modification adjusts the contribution of clean and noisy samples to the loss (Reed et al. 2014; Zheng et al. 2020). Third, sample selection methods concentrate on choosing clean samples from noisy ones. For example, Arpit et al. (2017) demonstrates the memorization effect of networks that always prefer to learn simple samples before fitting noisy data. Motivated by the memorization effect, Arazo et al. (2019) adopts a two-component Gaussian Mixture Model (GMM) to fit per-sample loss and treats the samples with minor loss as clean samples. To transfer the above noisy label learning technique from the classification problem to the cross-matching problem, Huang et al. (2021) proposes noisy correspondence learning. Amrani et al. (2021) use density of similarity to estimate the
noise probability. Thomas and Kovashka (2022) apply semantic neighborhood discrepancy and diversity to capture the degree of abstractness of an image-text pair. Different from them, NLIP introduces a new noise-adaptive image-text contrastive loss that harmonizes the cross-modal alignment by considering varying noise probabilities of different pairs and also rectifies the noisy samples via a concept-guided captioner. NLIP would be one of the early attempts that provide effective and efficient schemes within a large-scale image-text pre-training framework. It can be coupled with any VLP models to improve their robustness.

**Method**

We proposed Noise-robust Language-Image Pre-training framework (NLIP), a new VLP framework to learn from noisy image-text pairs. In this section, we first introduce the overall model architecture of NLIP. Then we present the model details in two noisy learning schemes respectively, including the noise-harmonization scheme to harmonize the cross-modal alignment with noise-adaptive regularization and the noise-completion scheme to enrich the missing object information of text.

**Basic Notations.** We use \( D = \{X, Y\} \) to denote the image-text dataset with the images \( X = \{x_i\}_{i=1}^N \) and texts \( Y = \{y_i\}_{i=1}^N \), where \( N \) denotes the total number of image-text pairs of the dataset. For vision modality, \( \mathcal{V}_e \) and \( \mathcal{V}_d \) denote vision encoder and vision decoder respectively. For language modality, \( \mathcal{T}_e \) denotes the text encoder. We denote the concept-conditioned cross-modal decoder by \( \mathcal{C}_d \).

**Overall Architecture**

Fig. 2 illustrates an overview of NLIP architecture for learning the high-quality cross-modal feature alignment. NLIP contains a visual encoder-decoder inspired by MAE (He et al. 2021) for reducing the computation cost and maintaining the high quality of visual feature representation, a text encoder encoding the texts enriched by extra auxiliary visual concepts and a concept-conditioned cross-modal decoder learning to synthesize semantic-consistent captions to complete noisy ones. For visual modality, we use Vision Transformer(ViT) (Dosovitskiy et al. 2020) that takes the concatenation of an extra [CLS] token embedding and linearly projected image patches as input and output the [CLS] token to represent the global image feature. Specifically, we randomly mask the patches and skip the mask token to reduce the computation cost. To enhance visual feature representation via self-supervised regularization, an MAE decoder is adopted to restore masked patches by Image Reconstruction (IR) loss \( \mathcal{L}_{IR} \):

\[
\mathcal{L}_{IR} = \sum_{i=1}^N \left( \frac{\mathcal{V}_e(x'_i)}{\|\mathcal{V}_e(x'_i)\|} - \frac{x_i}{\|x_i\|} \right)^2,
\]

where \( \| \cdot \| \) denotes the normalization, and \( x'_i \) represents masked patches. As for the language modality, we exploit an encoder-decoder structure to obtain the generation capability and synthesize enriched captions. We first retrieve the visual concepts (i.e., names of existing objects) for each input image from a large corpus via a pre-trained model.
The visual concepts concatenated with corresponding input texts are encoded by text encoder. Then a concept-conditioned cross-modal decoder is trained with the Language Modeling (LM) loss $L_{LM}$ to generate a more detailed caption for each image guided by the visual concepts. For the cross-modal alignment, the Noise-adaptive Image-Text Contrastive (NITC) loss $L_{NITC}$ is conducted to not only encourage the positive pair representations to get closer contrast to the negative pairs but also introduce the noise-adaptive label smoothing as an instance-aware regularization for avoiding severe bias to the noisy data. Therefore, the overall loss can be written as:

$$L = L_{IR} + \alpha \cdot L_{LM} + \beta \cdot L_{NITC}.$$ (2)

where $\alpha$ and $\beta$ denote the weighting factors.

### Noise Harmonization

To avoid over-fitting to the noisy image-text pairs, NLIP introduces the noise harmonization scheme by learning to harmonize the cross-modal alignments and adopts noise-adaptive regularization for each pair based on the estimated noisy probability.

#### Preliminaries.

To align between two different modalities, current vision-language pre-training models (Radford et al. 2021) adopt the Image-Text Contrastive (ITC) loss, to encourage positive image-text pairs $\{x_i, y_j\}_{i=j}$ aligned in the same feature space while in contrast to the negative pairs $\{x_{i}, y_{j}\}_{i \neq j}$. The normalized features from the image encoder and text encoder are denoted as $\mathcal{V}_e(x_i)$ and $T_e(y_j)$. We first calculate the per-sample image-text similarity $s^B \in \mathbb{R}^{B \times B}$ and text-to-image similarity $s^T$ in a batch as:

$$s_{i,j}^B = s_{i,j}^T = \mathcal{V}_e(x_i)'^T T_e(y_j).$$ (3)

where $B$ denotes the batch size. Then the Image-Text Contrastive loss $L_{ITC}$ can be written as the average of image-to-text and text-to-image contrastive loss:

$$L_{ITC} = \frac{1}{2B} \sum_{i=1}^B (L_i^T + L_i^y).$$ (4)

$$L_i^T = L_i^T(x_i, \{y_j\}_{j=1}^B) = -\log \frac{\exp(s^x_{i,j})}{\sum_j \exp(s^x_{i,j})},$$ (5)

$$L_i^y = L_i^y(y_i, \{x_j\}_{j=1}^B) = -\log \frac{\exp(s^y_{i,j})}{\sum_j \exp(s^y_{i,j})}.\quad (6)$$ (6)

However, existing ITC loss forces models to align the feature of each image-text pair without considering the situation that many of them are noisy. Directly pre-training with these samples may degrade the model performance.

### Noise-adaptive Image-Text Contrastive Loss.

We further propose a Noise-adaptive Image-Text Contrastive (NITC) loss $L_{NITC}$ to harmonize the cross-modal alignments with varying degrees according to its noisy probability. We first calculate the noisy probability of each image-text pair, which indicates the image and text in this pair are not semantically matched, according to the memorization effect (Arpit et al. 2017; Zhang et al. 2021a). Specifically, the cross-modal transformer tends to fit the easy (i.e., clean) samples first and then the noisy ones. Therefore, we adopt a two-component Gaussian Mixture Model (GMM) (Permuter et al. 2006) to fit the per-sample ITC loss. Specifically, we consider the probability predicted by the higher mean component as noisy probability $\varepsilon_i$ of $i$-th image-text pair, inspired by (Huang et al. 2021; Arazo et al. 2019):

$$p(L_{ITC}(x_i, y_i)|\theta) = \sum_{m=1}^2 \gamma_m \phi(L_{ITC}(x_i, y_i)|m),$$ (7)

$$\varepsilon_i = p(\mu_n)p(L_{ITC}(x_i, y_i)|\mu_h)/p(L_{ITC}(x_i, y_i)).$$ (8)

where $\gamma_m$ denotes the mixture coefficient, $\phi(m)$ is the probability density of the $m$-th GMM component, $\theta$ represents the parameters of GMM, and $\mu_h$ denotes the component with a higher mean.

Then we directly regularize the ground-truth alignment label with various degrees considering its noisy probability $\varepsilon_i$. Lower regularization is adopted for the clean samples (i.e., with low $\varepsilon_i$) to learn the alignment, while the higher regularization is adopted for noisy samples (i.e., with high $\varepsilon_i$) to avoid over-fitting the noise. In detail, inspired by the label-smoothing (Szegedy et al. 2016), we regularize the ground-truth image-to-text and text-to-image alignment label with different smoothing rates $W = \{w_i\}_{i=1}^{\text{NITC}}$, which is linearly associated with the noisy probability of each sample $\{w_i = \lambda \varepsilon_i, w_i \in [0, \lambda]\}$. $\lambda$ denotes the hyper-parameter to control the range of smooth rate. Then the Noise-adaptive Image-Text Contrastive loss $L_{NITC}$ is defined as:

$$L_{NITC} = \frac{1}{2B} \sum_{i=1}^B (\hat{L}_i^T + \hat{L}_i^y).$$ (9)

$$\hat{L}_i^T = -\log \frac{(1-w_i) \exp(s^x_{i,i})}{\sum_{j \neq i} (1-w_j) \exp(s^x_{i,j}) + \frac{w_i}{B-1} \sum_{j \neq i} \exp(s^y_{i,j})},$$ (10)

$$\hat{L}_i^y = -\log \frac{(1-w_i) \exp(s^y_{i,i})}{\sum_{j} (1-w_j) \exp(s^y_{i,j}) + \frac{w_i}{B-1} \sum_{j} \exp(s^y_{i,j})},$$ (11)

### Noise Completion

Apart from adopting the above instance-ware regularization on the noisy pairs, NLIP also introduces the noise completion scheme to enrich the missing object information of text since the captions from the web are naturally incomplete. Especially, NLIP injects a concept-conditioned cross-modal decoder to obtain semantic-consistent synthetic captions to complete noisy ones, which uses the retrieved visual concepts (i.e., names of existing objects) for the corresponding image to guide captioning generation.

#### Visual Concept.

Although the image-text data can be easily crawled from the web, the texts usually contain much noise, including missing details of the image and carrying unrelated contents to the image (Li et al. 2022a). To better address the problem of image-text misalignment, we introduce the visual concepts $q_c$ as auxiliary inputs to provide the prior information of existing objects for each image. We first construct a large visual concept vocabulary $Q$ via parsing the various concept nouns from the web-collected corpus. Then we retrieve the words of top-k similarity with image $x_i$ as
visual concepts $q_i \in Q$ based on a pre-trained VLM for that image. The similarity $\text{sim}(x_i, Q)$ between the input image $x_i$ and the nouns in $Q$ is calculated by

$$\text{sim}(x_i, Q) = \langle V_e(x) \cdot T_e([p, q]) \rangle.$$  

where $p$ denotes the pre-defined text prompt that is aggregated with the visual concepts to narrow down the gap with natural language (Radford et al. 2021). Based on the retrieved visual concepts $q_i$, NLIP uses an additional concept-conditioned cross-modal decoder (shown in Fig. 2) to synthesize new texts $Y'$ to replace the original texts $Y$ in noisy image-text pairs. Specifically, the cross-modal decoder is optimized by recovering the masked texts $y^{\text{mask}}$ with an autoregressive (i.e., language modeling) loss:

$$\mathcal{L}_{\text{LM}} = -\mathbb{E}_{(x,y) \sim D} \log p(y_{t \text{<} i}[C_d(y_{t \text{<} i})[V_e(x), T_e(z)])].$$  

where $[\cdot]$ denotes the concatenation operation, $t$ denotes the word index of text $y$, $z$ denotes the concatenation $[p, q, y^{\text{mask}}]$. Note that we omit index $i$ here.

**Pre-training procedure**

As shown in Fig. 3, we divide the whole pre-training paradigm of NLIP into three steps: noisy-aware pre-training, captioning and conception-enhanced pre-training. At noisy-aware pre-training stage, we adopt the noisy-adaptive regularization to pre-train NLIP. At captioning stage, we use captioning data to train concept-conditioned cross-modal decoder and generate synthetic captions for web images. At conception-enhanced pre-training stage, we select training captions by noisy probabilities and fine-tune NLIP.

![Figure 3: Illustration of NLIP procedure. The whole pre-training contains three stages: noisy-aware pre-training, captioning and conception-enhanced pre-training. At noisy-aware pre-training, we adopt the noisy-adaptive regularization to pre-train NLIP. At captioning stage, we use captioning data to train concept-conditioned cross-modal decoder and generate synthetic captions for web images. At conception-enhanced pre-training stage, we select training captions by noisy probabilities and fine-tune NLIP.](image_url)
Table 1: Top-1 accuracy(%) of zero-shot image classification and linear probing image classification tasks on 12 datasets when pre-training on YFCC26M.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>Caltech101</th>
<th>StanfordCars</th>
<th>Flowers102</th>
<th>Food101</th>
<th>SUN397</th>
<th>DTD</th>
<th>Aircrafts</th>
<th>OxfordPets</th>
<th>EuroSAT</th>
<th>ImageNet</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>74.8</td>
<td>44.1</td>
<td>64.5</td>
<td>3.7</td>
<td>51.4</td>
<td>45.1</td>
<td>43.7</td>
<td>14.5</td>
<td>4.3</td>
<td>22.9</td>
<td>23.0</td>
<td>34.8</td>
<td>35.6</td>
</tr>
<tr>
<td>FILIP</td>
<td>83.6</td>
<td>51.7</td>
<td>73.6</td>
<td>7.8</td>
<td>60.5</td>
<td>55.9</td>
<td>47.9</td>
<td>18.8</td>
<td>8.0</td>
<td>29.9</td>
<td>29.5</td>
<td>41.4</td>
<td>42.4</td>
</tr>
<tr>
<td>NLIP</td>
<td>74.0</td>
<td>47.4</td>
<td>75.1</td>
<td>6.8</td>
<td>58.9</td>
<td>53.8</td>
<td>55.4</td>
<td>32.3</td>
<td>8.9</td>
<td>36.8</td>
<td>35.4</td>
<td>42.4</td>
<td>43.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backbone</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>Caltech101</th>
<th>StanfordCars</th>
<th>Flowers102</th>
<th>Food101</th>
<th>SUN397</th>
<th>DTD</th>
<th>Aircrafts</th>
<th>OxfordPets</th>
<th>EuroSAT</th>
<th>ImageNet</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>75.3</td>
<td>42.4</td>
<td>69.5</td>
<td>3.9</td>
<td>54.8</td>
<td>51.1</td>
<td>46.6</td>
<td>18.6</td>
<td>3.9</td>
<td>21.7</td>
<td>20.5</td>
<td>39.2</td>
<td>37.3</td>
</tr>
<tr>
<td>FILIP</td>
<td>83.8</td>
<td>51.2</td>
<td>76.1</td>
<td>8.9</td>
<td>62.8</td>
<td>63.5</td>
<td>52.5</td>
<td>21.8</td>
<td>10.2</td>
<td>36.7</td>
<td>24.9</td>
<td>46.7</td>
<td>44.9</td>
</tr>
<tr>
<td>NLIP</td>
<td>81.9</td>
<td>47.5</td>
<td>79.5</td>
<td>7.8</td>
<td>59.2</td>
<td>58.7</td>
<td>32.9</td>
<td>7.5</td>
<td>39.2</td>
<td>33.9</td>
<td>47.4</td>
<td>45.9</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Table 2: Results of zero-shot image-to-text and text-to-image retrieval on Flickr30K. * means the model is fine-tuned on MSCOCO dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unicoder-VL</td>
<td>64.3</td>
<td>85.8</td>
<td>92.3</td>
<td>48.4</td>
<td>76.0</td>
<td>85.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageBERT</td>
<td>70.7</td>
<td>90.2</td>
<td>94.0</td>
<td>54.3</td>
<td>79.6</td>
<td>87.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNITER</td>
<td>80.7</td>
<td>95.7</td>
<td>98.0</td>
<td>66.2</td>
<td>88.4</td>
<td>92.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP(ViT-B/32)</td>
<td>46.4</td>
<td>75.4</td>
<td>84.1</td>
<td>29.8</td>
<td>56.1</td>
<td>67.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FILIP(ViT-B/32)</td>
<td>56.6</td>
<td>82.7</td>
<td>90.0</td>
<td>39.5</td>
<td>66.7</td>
<td>75.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLIP(ViT-B/32)</td>
<td>77.2</td>
<td>94.8</td>
<td>97.7</td>
<td>56.6</td>
<td>83.2</td>
<td>89.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP(ViT-B/16)</td>
<td>53.9</td>
<td>81.0</td>
<td>90.1</td>
<td>34.6</td>
<td>62.6</td>
<td>73.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP*(ViT-B/16)</td>
<td>73.5</td>
<td>92.6</td>
<td>96.2</td>
<td>54.1</td>
<td>81.9</td>
<td>89.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FILIP(ViT-B/16)</td>
<td>66.5</td>
<td>88.4</td>
<td>93.9</td>
<td>47.1</td>
<td>74.4</td>
<td>82.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLIP(ViT-B/16)</td>
<td>82.6</td>
<td>96.6</td>
<td>98.3</td>
<td>61.2</td>
<td>85.7</td>
<td>91.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 demonstrates the superior zero-shot transfer capability. These 12 classification datasets consist of CIFAR10 (Krizhevsky 2009), CIFAR100 (Krizhevsky 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2006), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014), SUN397 (Xiao et al. 2010), DTD (Cimpoi et al. 2014), Aircrafts (Maji et al. 2013), OxfordPets (Parkhi et al. 2012), EuroSAT (Helber et al. 2019), ImageNet (Russakovsky et al. 2015), covering a wide range of domains.

Zero-Shot Image Classification. Experimental results show that NLIP largely outperforms the corresponding baseline CLIP in terms of average top-1 accuracy over 12 datasets and achieves an improvement of 8.6%. In particular, NLIP surpasses CLIP on ImageNet over 8.2%. Besides, NLIP also obtains substantial performance gains in most individual datasets with images in different domains, demonstrating the effectiveness of proposed noise-harmonization and noise completion schemes. Compare to FILIP which learns the finer-grained alignment between image and text, NLIP with global image-text alignment achieves 1.0% average improvement over 12 datasets.

Linear Probing Image Classification. Table 1 demonstrates that NLIP achieves 76.9% on average top-1 accuracy over 12 downstream tasks, which surpasses FILIP and CLIP by 7.4% and 6.5%, respectively. NLIP with ViT-B/32 also outperforms FILIP and CLIP about 4.9% and 4.7%. The linear probing experiments demonstrate the robustness representation learned by NLIP.

Image-Text Retrieval

We evaluate NLIP on both zero-shot image-to-text retrieval (TR) and zero-shot text-to-image retrieval (IR) tasks on Flickr30K (Plummer et al. 2015). Then we also compare NLIP against the existing vision-language pre-training methods, including Unicoder-VL (Li et al. 2020), ImageBERT (Qi et al. 2020), UNITER (Chen et al. 2020). These
models are single-stream and employ an additional object detector to extract region features while NLIP only employs visual patch features for simplicity.

Table 2 demonstrates that NLIP achieves substantial improvement compared to CLIP pre-trained in YFCC26M. In image-to-text retrieval, NLIP outperforms CLIP by 28.7% in R@1. In text-to-image retrieval, NLIP is 26.6% higher than CLIP on R@1 and 7.1% higher than CLIP* fine-tuned on MSCOCO dataset. NLIP also achieves a 1.9% improvement over UNITER in R@1. As shown in Table 4, when only using YFCC26M-pretrained CLIP to retrieve visual concepts, our NLIP still beats CLIP and CLIP* over 23.2% and 3.6% on zero-shot image-to-text retrieval task, which demonstrates the superiority of the noise-robust learning in NLIP under the exact same pre-training data.

**Image Captioning**

We further evaluate the pre-trained NLIP on downstream image captioning task, which aims at generating the description of an image in natural language, on COCO Caption (Lin et al. 2014) dataset. We evaluate different methods on standard metrics for the captioning task, including BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). For fair comparison with other models, we follow BLIP (Li et al. 2022a) to initialize the visual encoder of NLIP from an ImageNet pre-trained ViT-B/16.

As shown in Table 3, NLIP achieves 40.3 in BLEU@4 and 135.2 in CIDEr, outperforming CLIP (Li et al. 2022a) by 1.9 in CIDEr. Note that BLIP is pre-trained with 5x more image-text pairs (129M v.s. 26M). NLIP with train-from-scratch image encoder still outperforms BLIP, according to the third row of Table 4. NLIP also beats other methods (e.g., SimVLM) pre-trained on large-scale datasets. Particularly, VinVL (Zhang et al. 2021b) requires an object detector pre-trained on 2.5M images with high resolution (800×1333) and full human-annotated bounding boxes.

**Ablation Studies**

**Effect of Noise Harmonization.** Table 4 ablates the effectiveness of our noise harmonization. By comparing with the last two rows, we can find that NLIP gains 1.2% and 2.5% improvement in image-to-text retrieval and text-to-image retrieval with noise harmonization, respectively, verifying that pre-training with NITC loss helps the model avoid overfitting on the mismatched image-text pairs.

**Effect of Noise Completion.** Table 4 shows that NLIP with the noise completion scheme can boost performance on all downstream tasks. We can observe the noise completion scheme helps boost the image caption task by over 1.6% on CIDEr and the text retrieval task by 10.4% on R@1. Besides, without the condition of visual concepts in noise completion, NLIP will drop 0.7% accuracy on zero-shot ImageNet classification and 1.1% R@1 on image retrieval of Flickr30K. Incorporating visual concepts into the cross-modal decoder further help enrich the synthetic caption with more information of existing objects and boost the performance in all downstream tasks, as shown in Table 4.

**Conclusion**

In this paper, we propose a new vision-language pre-training framework named NLIP to learn from the noisy image-text pairs crawled from the web. NLIP introduces two schemes, including noise-harmonization and noise-completion, to stabilize the pre-training and efficiently make full use of noisy pairs. In noise-harmonization scheme, NLIP adopts noise-adaptive regularization to harmonize the cross-modal alignments with varying degrees by considering the noise probability of each pair. And in noise-completion scheme, NLIP further introduces a concept-conditioned cross-modal decoder to obtain synthetic captions to complete noisy ones. Retrieved visual concepts are utilized as the auxiliary input for the cross-modal decoder to provide the prior information of existing objects. Experiments show that NLIP achieves significant performance gaps on several downstream tasks, including zero-shot classification, image-text retrieval and caption generation tasks. In the future, our NLIP can be easily injected into any cross-modal pre-training models and the proposed noisy-robust learning schemes can be beneficial for more downstream fine-grained tasks such as open-world object detection, segmentation, and image generation.

<table>
<thead>
<tr>
<th>Model</th>
<th># Pre-train Images</th>
<th>MSCOCO BLEU@4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder-Decoder</td>
<td>15M</td>
<td>-</td>
<td>110.9</td>
</tr>
<tr>
<td>BUTD</td>
<td>1.7M</td>
<td>36.4</td>
<td>120.1</td>
</tr>
<tr>
<td>VinVL</td>
<td>5.7M</td>
<td>38.2</td>
<td>129.3</td>
</tr>
<tr>
<td>VLP</td>
<td>3M</td>
<td>39.5</td>
<td>129.8</td>
</tr>
<tr>
<td>AoANet</td>
<td>1.7M</td>
<td>38.9</td>
<td>129.8</td>
</tr>
<tr>
<td>UNIMO&lt;sub&gt;base&lt;/sub&gt;</td>
<td>11.3M</td>
<td>38.8</td>
<td>124.4</td>
</tr>
<tr>
<td>SimVLM&lt;sub&gt;base&lt;/sub&gt;</td>
<td>1.8B</td>
<td>39.0</td>
<td>134.8</td>
</tr>
<tr>
<td>BLIP</td>
<td>129M</td>
<td>39.7</td>
<td>133.3</td>
</tr>
<tr>
<td><strong>NLIP</strong></td>
<td><strong>26M</strong></td>
<td><strong>40.3</strong></td>
<td><strong>135.2</strong></td>
</tr>
</tbody>
</table>

Table 3: Comparison with SoTA image captioning methods on COCO captioning benchmark. NLIP achieves the best performance even using a small-scale pre-training dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ImageNet Pre-train</th>
<th>COCO Pre-train</th>
<th>Flickr30K Pre-train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td>ZS-CLS Top-1 R@1 R@10</td>
<td>BLEU CIDEr image-to-text text-to-image R@1 R@10</td>
<td></td>
</tr>
<tr>
<td>CLIP</td>
<td>-</td>
<td>-</td>
<td>53.9 90.1 34.6 73.6</td>
</tr>
<tr>
<td><strong>NLIP</strong></td>
<td><strong>43.0</strong></td>
<td><strong>39.0</strong></td>
<td><strong>130.6</strong></td>
</tr>
<tr>
<td>w/o VC</td>
<td>46.7</td>
<td>39.6</td>
<td>132.8</td>
</tr>
<tr>
<td>w/o NC</td>
<td>47.0</td>
<td>39.6</td>
<td>132.4</td>
</tr>
<tr>
<td>w/o NH</td>
<td>46.7</td>
<td>39.6</td>
<td>131.5</td>
</tr>
</tbody>
</table>

Table 4: Ablation studies of all components on zero-shot classification, image-text retrieval and image caption. We denote using condition of visual concepts in noise completion as “VC”, noise completion as “NC”, and noise harmonization as “NH”. Note that removing the noise completion scheme degrades the performance severely. † denotes using the YFCC26M-pretrained CLIP to retrieve visual concepts.
Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant No. 22lgqj38. We gratefully acknowledge the support of MindSpore\(^1\), CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References


Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerrig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML.


\(^1\)https://www.mindspore.cn/


