NLIP: Noise-Robust Language-Image Pre-training

Runhui Huang[1], **Yanxin Long**[1], **Jianhua Han**[2], **Hang Xu**[2], **Xiwen Liang**[1], **Chunjing Xu**[2], **Xiaodan Liang**[1 *]

[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 successes highly rely 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 **N**oise-robust **L**anguage-**I**mage **P**re-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 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

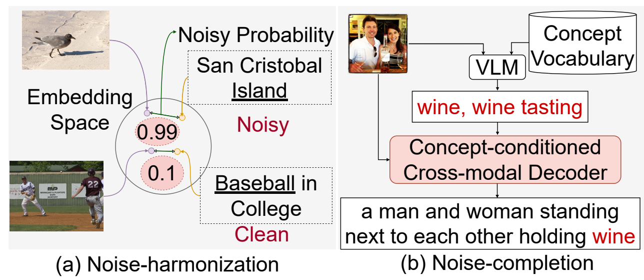

Figure 1: Illustration of two proposed schemes. (a) *Noise-harmonization*: NLIP estimates the noise probability of each image-text pair and enforces the pairs with larger noise probability to have fewer similarities in embedding space. (b) *Noise-completion*: NLIP generates enriched descriptions via a concept-conditioned captioner by taking visual concepts retrieved from a vocabulary as auxiliary inputs.

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

*Corresponding author

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 **N**oise-robust **L**anguage-**I**mage **P**re-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

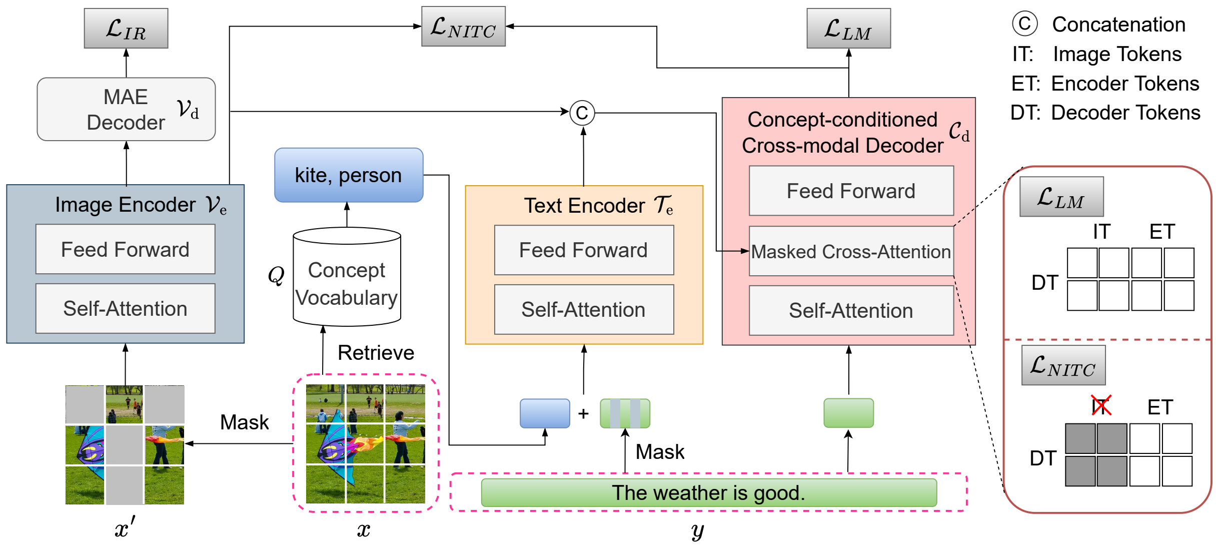

Figure 2: Overview of the proposed NLIP architecture. NLIP consists of an image encoder $\mathcal{V}_e$, text encoder $\mathcal{T}_e$, cross-modal decoder $\mathcal{C}_d$ and MAE decoder $\mathcal{V}_d$. During training, given an input image x, it feeds the randomly masked visual patches into an image encoder and the MAE decoder learns to reconstruct them via $\mathcal{L}_{IR}$. The correlated concepts are also retrieved from a vocabulary for each image and then concatenated with the text y as inputs of the text encoder. The concept-conditioned cross-modal decoder is fed with image features, concept-conditioned text features and text embedding, and optimized via $\mathcal{L}_{LM}$. The noise-adaptive image-text contrastive loss $\mathcal{L}_{NITC}$ is adopted to learn cross-modal alignment by considering varying noise probabilities. Note that the concept-conditioned cross-modal decoder does not utilize image tokens as input for $\mathcal{L}_{NITC}$ to avoid information leakage while does for $\mathcal{L}_{LM}$. Omit the index i here.

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 the 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 **N**oise-robust **L**anguage-**I**mage **P**re-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 $\mathcal{L}_{IR}$:

$$\mathcal{L}_{IR} = \sum_{i=1}^{N} \left(\frac{\mathcal{V}_e(x_i')}{\|\mathcal{V}_e(x_i')\|} - \frac{x_i}{\|x_i\|} \right)^2. \tag{1}$$

where $\|\cdot\|$ denotes the normalization, and x' 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 $\mathcal{L}_{LM}$ 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 $\mathcal{L}_{NITC}$ 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:

$$\mathcal{L} = \mathcal{L}_{IR} + \alpha \cdot \mathcal{L}_{LM} + \beta \cdot \mathcal{L}_{NITC}. \tag{2}$$

where α and β 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 $\mathcal{T}_e(y_i)$. We first calculate the per-sample image-to-text similarity $s^y \in \mathbb{R}^{B\times B}$ and text-to-image similarity s^x in a batch as:

$$s^y_{i,j} = s^x_{i,j} = \mathcal{V}_e(x_i)^\top \mathcal{T}_e(y_j). \tag{3}$$

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

$$\mathcal{L}_{ITC} = \frac{1}{2B}\sum_{i=1}^{B}(\mathcal{L}^x_i + \mathcal{L}^y_i), \tag{4}$$

$$\mathcal{L}^x_i = \mathcal{L}^x_i(x_i, \{y_j\}^B_{j=1}) = -\log\frac{\exp(s^x_{i,i})}{\sum_j \exp(s^x_{i,j})}, \tag{5}$$

$$\mathcal{L}^y_i = \mathcal{L}^y_i(y_i, \{x_j\}^B_{j=1}) = -\log\frac{\exp(s^y_{i,i})}{\sum_j \exp(s^y_{i,j})}. \tag{6}$$

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 $\mathcal{L}_{NITC}$ 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 ε_i of i-th image-text pair, inspired by (Huang et al. 2021; Arazo et al. 2019):

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

$$\varepsilon_i = p(\mu_h)p(\mathcal{L}_{ITC}(x_i, y_i)|\mu_h)/p(\mathcal{L}_{ITC}(x_i, y_i)). \tag{8}$$

where γ_m denotes the mixture coefficient, $\phi(\cdot|m)$ is the probability density of the m-th GMM component, θ represents the parameters of GMM, and μ_h denotes the component with a higher mean.

Then we directly regularize the ground-truth alignment label with various degrees considering its noisy probability ε_i. Lower regularization is adopted for the clean samples (*i.e.*, with low ε_i) to learn the alignment, while the higher regularization is adopted for noisy samples (*i.e.*, with high ε_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 $W = \{w_i\}^N_{i=1}$, which is linearly associated with the noisy probability of each sample $\{w_i = \lambda\varepsilon_i,\ w_i \in [0, \lambda]\}$. λ denotes the hyper-parameter to control the range of smooth rate. Then the Noise-adaptive Image-Text Contrastive loss $\mathcal{L}_{NITC}$ is defined as:

$$\mathcal{L}_{NITC} = \frac{1}{2B}\sum_{i=1}^{B}(\hat{\mathcal{L}}^x_i + \hat{\mathcal{L}}^y_i), \tag{9}$$

$$\hat{\mathcal{L}}^x_i = -\log\frac{(1-w_i)\exp(s^x_{i,i})}{(1-w_i)\exp(s^x_{i,i}) + \frac{w_i}{B-1}\sum\limits_{i\neq j}\exp(s^x_{i,j})}, \tag{10}$$

$$\hat{\mathcal{L}}^y_i = -\log\frac{(1-w_i)\exp(s^y_{i,i})}{(1-w_i)\exp(s^y_{i,i}) + \frac{w_i}{B-1}\sum\limits_{i\neq j}\exp(s^y_{i,j})}. \tag{11}$$

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_v 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

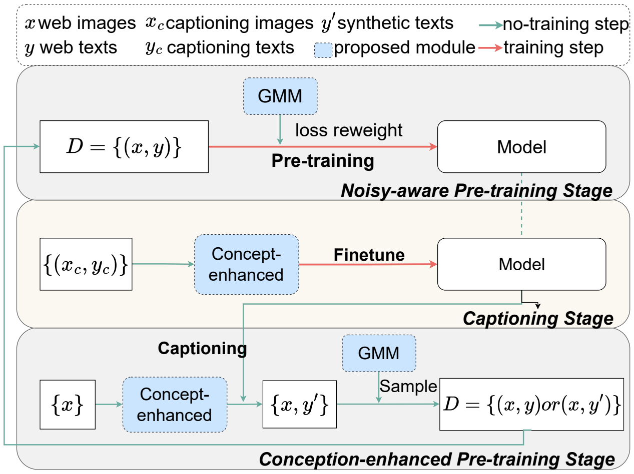

Figure 3: Illustration of NLIP procedure. The whole pre-training contains three stages: *noisy-aware pre-training*, *captioning* and *conception-enhanced pre-training*. At *noisy-aware pre-training* stage, we adopt the noisy-adaptive regularization to pre-train NLIP. At *captioning* stage, we use captioning 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.

visual concepts $q_i \in Q$ based on a pre-trained VLM for that image. The similarity $sim(x_i, Q)$ between the input image x_i and the nouns in Q is calculated by

$$sim(x_i, Q) = \langle \mathcal{V}_{\mathrm{e}}(x) \cdot \mathcal{T}_{\mathrm{e}}([\mathrm{p}, Q]) \rangle. \tag{12}$$

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^m with an autoregressive (*i.e.*, language modeling) loss:

$$\mathcal{L}_{LM} = -\mathbb{E}_{(x,y)\sim D} \log p(y_t | \mathcal{C}_{\mathrm{d}}(y_{\tau<t}, [\mathcal{V}_{\mathrm{e}}(x), \mathcal{T}_{\mathrm{e}}(z)])). \tag{13}$$

where $[\cdot]$ denotes the concatenation operation, t denotes the word index of text y, z denotes the concatenation $[\mathrm{p}, q, y^m]$. 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 first warm up the NLIP architecture with E_e epochs under the supervision of $\mathcal{L}_{IR}$, $\mathcal{L}_{LM}$ and $\mathcal{L}_{ITC}$. Then we estimate the noisy probability ε_i of the i-th image-text pair based on the $\mathcal{L}_{ITC}$ and adopt the noisy-adaptive regularization by replacing the $\mathcal{L}_{ITC}$ with $\mathcal{L}_{NITC}$ in the following E_t epochs. At *captioning* stage, to obtain better generation ability, we further fine-tune the captioner, which includes the image encoder $\mathcal{V}_{\mathrm{e}}$, text encoder $\mathcal{T}_{\mathrm{e}}$ and cross-modal decoder $\mathcal{C}_{\mathrm{d}}$, on captioning dataset COCO Captions (Lin et al. 2014) and generates new texts $Y' = \{y'_i\}_{i=1}^N$ for each image-text pair. Finally, at *conception-enhanced pre-training* stage, we fine-tune NLIP with E_f epochs with the revised image-text pairs D', where each text y_i of the i-th pair in original dataset D is replaced by the synthetic text y'_i randomly with sampling rate same as the noisy probability ε_i.

Experiments

Experimental Settings

Model Architecture. We adopt the ViT-B/16 and ViT-B/32 as our visual encoder architecture. Unless specified, NLIP uses ViT-B/16 as the visual encoder. The text encoder and concept-conditioned cross-modal decoder are initialized from $\text{BART}_{\text{base}}$ (Lewis et al. 2020) and the MAE decoder only has 4 transformer blocks with 64-d head.

Training Details. We pre-train our NLIP on 32 Nvidia V100 for 50 epochs with 6144 batch size. LAMB (You et al. 2020) optimizer is adopted with a weight decay of 0.05. The base learning rate is set to 0.003 and the scaling rule keeps the same with Yao et al. (2021). The learning rate is linearly warmed up in the first five epochs and then gets decayed by the cosine learning rate schedule (Loshchilov and Hutter 2016). We pre-train NLIP on a 26M subset of YFCC100M named YFCC26M, and the filtering rules follow FILIP (Yao et al. 2021). During the pre-training, the images are randomly cropped between 50% and 100% of the original size and then resized to 224 × 224 resolution. The visual encoder applies 50% masking ratio. When conducting downstream tasks (*e.g.*, image captioning), the image resolution is resized to 384 × 384 and we don't mask any image patches. The training epochs E_e, E_t and E_f in different stages are set as 5, 45 and 20, respectively. The weighting factor α and β are both 1 and λ in $\mathcal{L}_{NITC}$ is 0.5. During *captioning* stage, following BLIP (Li et al. 2022a), we fine-tune NLIP on COCO (Lin et al. 2014)'s Karpathy train split (Karpathy and Fei-Fei 2015) to generate high-quality captions.

Visual Concept Vocabulary. The visual concept vocabulary Q is built by parsing the nouns from text corpus via spaCy toolkit and filtering nouns that appear less than 5 times. The source corpus includes YFCC100M (Thomee et al. 2016), OpenWebText (Gokaslan and Cohen 2019), WordNet of NLTK (Natural Language Toolkit) (Loper and Bird 2002) and the most-frequent n-gram collected from web.

After collecting, the visual concept vocabulary Q contains about 151k unique nouns. We use a pre-trained $\text{FILIP}_{\text{large}}$ (Yao et al. 2021) to retrieve visual concepts for each image. Unless specified, NLIP uses $\text{FILIP}_{\text{large}}$ to retrieve visual concepts. More ablation studies about the effect of utilizing different pre-trained VLMs (e.g. YFCC26m-pretrained CLIP-ViT-L/16) are shown in the ablation study.

Image Classification

We evaluate our NLIP on the zero-shot image classification and linear probing task on 12 downstream classifica-

	Backbone	CIFAR10	CIFAR100	Caltech101	StanfordCars	Flowers102	Food101	SUN397	DTD	Aircrafts	OxfordPets	EuroSAT	**ImageNet**	**Average**
Zero-Shot Image Classification														
CLIP	ViT-B/32	74.8	44.1	64.5	3.7	51.4	45.1	43.7	14.5	4.3	22.9	23.0	34.8	35.6
FILIP		**83.6**	**51.7**	73.6	**7.8**	**60.5**	**55.9**	47.9	18.8	8.0	29.9	29.5	41.4	42.4
NLIP		74.0	47.4	**75.1**	6.8	58.9	53.8	**55.4**	**32.3**	**8.9**	**36.8**	**35.4**	**42.4**	**43.9**
CLIP	ViT-B/16	75.3	42.4	69.5	3.9	54.8	51.1	46.6	18.6	3.9	21.7	20.5	39.2	37.3
FILIP		**83.8**	**51.2**	76.1	**8.9**	**62.8**	**63.5**	52.5	21.8	**10.2**	36.7	24.9	46.7	44.9
NLIP		81.9	47.5	**79.5**	7.8	54.0	59.2	**58.7**	**32.9**	7.5	**39.2**	**33.9**	**47.4**	**45.9**
Linear Probing Image Classification														
CLIP	ViT-B/32	90.4	69.7	84.7	23.8	91.5	70.7	66.3	66.1	32.7	61.0	96.0	60.3	67.8
FILIP		90.5	69.5	88.2	30.0	90.9	69.2	67.6	66.0	31.3	56.0	93.4	58.8	67.6
NLIP		**90.9**	**73.4**	**89.2**	**34.1**	**95.6**	**76.9**	**71.9**	**71.3**	**39.8**	**62.5**	**96.8**	**67.1**	**72.5**
CLIP	ViT-B/16	90.5	71.1	86.6	29.4	92.8	78.4	67.7	66.2	37.2	66.0	94.3	65.0	70.4
FILIP		90.6	67.4	88.6	32.8	93.7	71.8	69.8	68.5	35.7	59.4	93.7	62.3	69.5
NLIP		**92.8**	**74.2**	**90.4**	**41.2**	**97.5**	**85.0**	**75.9**	**74.3**	**43.4**	**79.2**	**96.8**	**71.8**	**76.9**

Table 1: Top-1 accuracy(%) of zero-shot image classification and linear probing image classification tasks on 12 datasets when pre-training on YFCC26M.

tion datasets as in Table 1, demonstrating 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. Note that the linear probing task only trains a randomly initialized linear classifier with a pre-trained frozen image encoder on the downstream datasets. We compare with other vision-language pre-training methods, including FILIP with the token reduction layer (Yao et al. 2021; Gu et al. 2022) and CLIP (Radford et al. 2021) under the same dataset (*i.e.*, YFCC26M) and the same evaluation settings in (Radford et al. 2021). For fair comparison, we pre-train CLIP with the same augmentation strategies as ours. We ensemble all prompts by averaging the text embeddings for each class across the prompt templates as in (Radford et al. 2021).

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.

	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL	64.3	85.8	92.3	48.4	76.0	85.2
ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5
UNITER	80.7	95.7	98.0	66.2	88.4	92.9
CLIP(ViT-B/32)	46.4	75.4	84.1	29.8	56.1	67.8
FILIP(ViT-B/32)	56.6	82.7	90.0	39.5	66.7	75.8
NLIP(ViT-B/32)	**77.2**	**94.8**	**97.7**	**56.6**	**83.2**	**89.8**
CLIP(ViT-B/16)	53.9	81.0	90.1	34.6	62.6	73.6
CLIP*(ViT-B/16)	73.5	92.6	96.2	54.1	81.9	89.8
FILIP(ViT-B/16)	66.5	88.4	93.9	47.1	74.4	82.5
NLIP(ViT-B/16)	**82.6**	**96.6**	**98.3**	**61.2**	**85.7**	**91.7**

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.

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

Model	# Pre-train Images	MSCOCO BLEU@4	CIDEr
Encoder-Decoder	15M	-	110.9
BUTD	1.7M	36.4	120.1
VinVL	5.7M	38.2	129.3
VLP	3M	39.5	129.8
AoANet	1.7M	38.9	129.8
$\text{UNIMO}_{\text{base}}$	11.3M	38.8	124.4
$\text{SimVLM}_{\text{base}}$	1.8B	39.0	134.8
BLIP	129M	39.7	133.3
NLIP	26M	**40.3**	**135.2**

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.

Dataset	ImageNet	COCO		Flickr30K			
Task	ZS-CLS	BLEU	CIDEr	image-to-text		text-to-image	
Metric	Top-1			R@1	R@10	R@1	R@10
CLIP	39.2	-	-	53.9	90.1	34.6	73.6
NLIP†	43.0	39.0	130.6	77.1	98.2	63.9	92.5
NLIP	**47.4**	**39.9**	**134.0**	**82.6**	98.3	**61.2**	**91.7**
w/o VC	46.7	39.6	132.8	82.2	**98.6**	60.1	91.6
w/o NC	47.0	39.6	132.4	72.2	96.1	49.6	84.2
w/o NH	46.7	39.6	131.5	71.0	95.6	47.1	82.0

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.

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 BLIP (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.

Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant No. 22lgqb38. We gratefully acknowledge the support of MindSpore[1], CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

Amrani, E.; Ben-Ari, R.; Rotman, D.; and Bronstein, A. 2021. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6644–6652.

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.; and McGuinness, K. 2019. Unsupervised label noise modeling and loss correction. In *ICML*, 312–321. PMLR.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, 233–242. PMLR.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101–mining discriminative components with random forests. In *ECCV*, 446–461. Springer.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *IEEE/CVF Conference on CVPR*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

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

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

Gokaslan, A.; and Cohen, V. 2019. Openwebtext corpus.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Niu, M.; Xu, H.; Liang, X.; Zhang, W.; Jiang, X.; and Xu, C. 2022. Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework. arXiv:2202.06767.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

[1]https://www.mindspore.cn/

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with Noisy Correspondence for Cross-modal Matching. *Advances in Neural Information Processing Systems*, 34.

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

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on CVPR*, 3128–3137.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11336–11344.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022b. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations*.

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

Loper, E.; and Bird, S. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. *ICLR 2017 (5th International Conference on Learning Representations)*.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *ICML*, 6448–6458. PMLR.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2021. SLIP: Self-supervision meets Language-Image Pre-training. *arXiv preprint arXiv:2112.12750*.

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

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

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

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Permuter, H.; Francos, J.; Jermyn, I.; et al. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4): 695–706.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE international conference on computer vision*, 2641–2649.

Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

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

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115(3): 211–252.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2020. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*.

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

Thomas, C.; and Kovashka, A. 2022. Emphasizing Complementary Samples for Non-Literal Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4632–4641.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on CVPR*, 4566–4575.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

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

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

You, H.; Zhou, L.; Xiao, B.; Codella, N.; Cheng, Y.; Xu, R.; Chang, S.-F.; and Yuan, L. 2022. Learning Visual Representation from Modality-Shared Contrastive Language-Image Pre-training. *arXiv preprint arXiv:2207.12661*.

You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; and Hsieh, C.-J. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *ICLR*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A New Foundation Model for Computer Vision. *arXiv preprint arXiv:2111.11432*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021a. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

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

Zheng, S.; Wu, P.; Goswami, A.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, 11447–11457. PMLR.