

ViLTA: Enhancing Vision-Language Pre-training through Textual Augmentation

Weihan Wang*

Zhen Yang*

Bin Xu[†]

Juanzi Li

Yankui Sun

Tsinghua University

{wangwh21, yangz21}@mails.tsinghua.edu.cn

{xubin, lijuanzi, syk}@tsinghua.edu.cn

Abstract

Vision-language pre-training (VLP) methods are blossoming recently, and its crucial goal is to jointly learn visual and textual features via a transformer-based architecture, demonstrating promising improvements on a variety of vision-language tasks. Prior arts usually focus on how to align visual and textual features, but strategies for improving the robustness of model and speeding up model convergence are left insufficiently explored.

In this paper, we propose a novel method ViLTA, comprising of two components to further facilitate the model to learn fine-grained representations among image-text pairs. For Masked Language Modeling (MLM), we propose a cross-distillation method to generate soft labels to enhance the robustness of model, which alleviates the problem of treating synonyms of masked words as negative samples in one-hot labels. For Image-Text Matching (ITM), we leverage the current language encoder to synthesize hard negatives based on the context of language input, encouraging the model to learn high-quality representations by increasing the difficulty of the ITM task. By leveraging the above techniques, our ViLTA can achieve better performance on various vision-language tasks. Extensive experiments on benchmark datasets demonstrate that the effectiveness of ViLTA and its promising potential for vision-language pre-training.

1. Introduction

Recent advancements in Vision-Language Pre-training (VLP) have achieved significant improvements in a wide range of multimodal tasks, such as visual question answering (VQA) [4], image captioning [3], and image-text retrieval [35, 45]. The target of VLP generally starts with training a model on massive image-text pairs in a self-supervised way, which empowers a new paradigm for fine-tuning various downstream tasks. Most recent VLP mod-

els [44, 31, 6, 66, 5] usually utilize a transformer-based architecture [55] with some specific training techniques (e.g. image-text contrastive learning (ITC), masked language modeling (MLM), and image-text matching (ITM)) to align visual and textual information. These models have achieved outstanding performance on a variety of multimodal benchmarks, which further advances the field of multimodal representation learning. However, the above VLP models also suffer from two vital limitations: (1) one-hot labels in MLM hinder the robustness of the model; (2) negative samples selection in ITM affects the model convergence and downstream performance.

In specific, the MLM task is designed to predict masked tokens in a given language input by utilizing both visual and textual features. However, compared to traditional MLM approaches in NLP [14], the MLM task in vision-language pre-training presents a unique limitation. VLP models use a pre-trained language model for secondary pre-training on image-text pairs, which may result in a loss of knowledge initially acquired from NLP datasets. Empirical results from previous studies [17] indicate that pre-training on multimodal datasets may lead to degraded performance on unimodal language tasks. Moreover, the presence of multiple candidate words to fill a masked position in an image-text pair can hinder the training of MLM. For instance, in the sentence "Two giraffes pace around their habitat", substituting "pace" with "walk" does not alter the sentence's meaning. Consequently, treating these words as negative samples in one-hot labels can impede MLM training.

As a popular pre-training task in VLP, the goal of ITM task is to distinguish positive and negative image-text pairs based on the learned representations. The common and straightforward way for negative selection is to randomly sample negatives for any given image-text pair. However, such a simple method can not provide more contributions for model convergence and result in sub-optimal performance. As a result, the model can easily achieve high accuracy in the first training epoch, usually above 99%. To further improve the model's ability to learn fine-grained representations, an effective method is to offer hard negative samples to make the pre-training task more challenging.

*Equal contribution.

[†]Corresponding author.

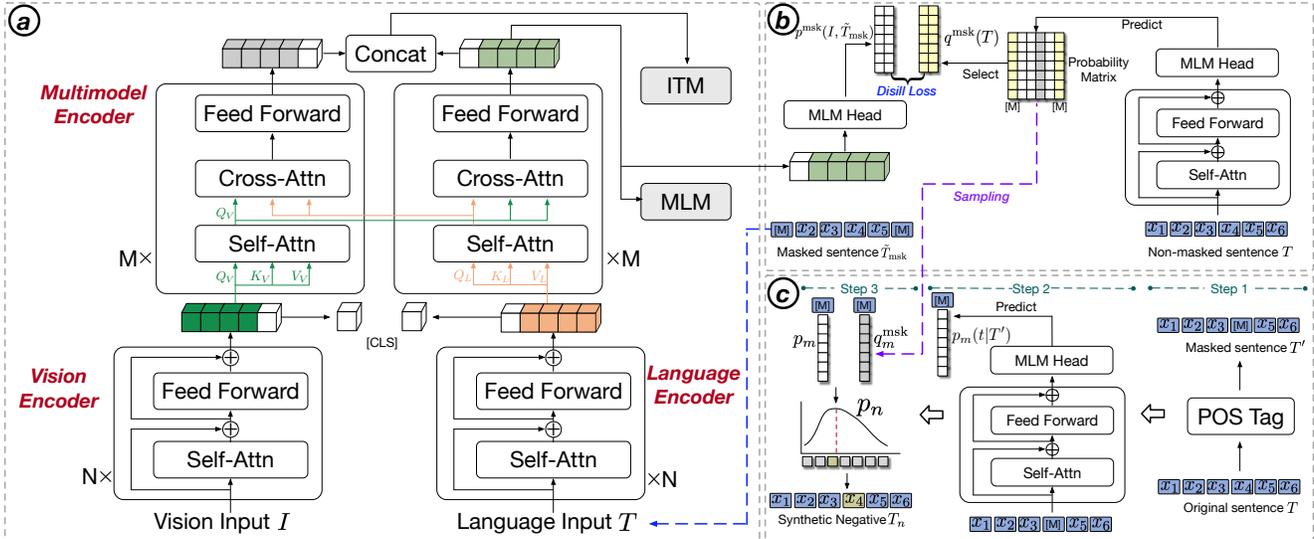


Figure 1. The overall architecture of ViLTA. The framework of ViLTA contains three components, including vision, language, and multi-modal encoders (Cf. (a)); soft labels obtained by the froze language encoder to enhance the robustness of model with noisy data in MLM (Cf. (b)); synthetic hard negatives generated by the current language encoder for ITM (Cf. (c)).

These hard negative samples are similar to positives and it is difficult for the model to distinguish them from positives. A growing body of research [49, 42, 21] illustrates that mining hard negatives can drastically alter the performance of multimodal models among various tasks, highlighting the significance of hard negative samples for enhancing the representation capabilities of the model. However, the existing attempts in mining hard negatives for vision-language pre-training only focus on sampling negatives in the discrete data space, ignoring the relationship among image-text pairs and the context of language input.

Present Work. To address the abovementioned issues, we propose a novel vision-language pre-training model named ViLTA, comprising two key components. For MLM, we propose a *cross-distillation* method to generate soft labels for improving the learning efficiency and boosting the robustness of the model. In specific, such a distillation method leverages the frozen language encoder to generate soft labels, which can be integrated into the original MLM task for joint training. In ITM, we propose to synthesize hard negatives based on the current language model by leveraging the context of language input, which is significantly different from previous works that select hard negatives from the raw data [31, 6]. By utilizing these two techniques, our proposed ViLTA can achieve better performance on a variety of downstream tasks, including visual question answering, visual entailment, visual reasoning, image-text retrieval, and image captioning. Extensive experimental results demonstrate the effectiveness of ViLTA. We summarize the contributions of this work as follows:

1) The proposed knowledge distillation method *cross-*

distillation generates soft labels to allow the model to better capture representations among image-text pairs, enabling the learning of the model more smooth and robust.

2) As opposed to sampling hard negatives from the raw data, we propose a strategy to synthesize hard negative samples based on the current language model, boosting the representation ability of the model by enhancing the difficulty of the ITM task.

3) By effectively integrating these two techniques, our ViLTA brings about outstanding performance improvements on various downstream tasks, demonstrating the superiority of ViLTA.

2. Related Work

Vision-Language Representation Learning. In prior research, several techniques have been utilized to integrate visual and language features for multimodal learning. One approach is to use pre-trained object detectors as feature extractors. ViLBERT [39] and LXMERT [53] employed co-attention for modality fusion, where two independent transformer modules were used for visual and language features, respectively. Alternatively, VisualBERT [32], VL-BERT [51], UNITER [11], OSCAR [34], VinVL [67], and VL-T5 [12] employed a merged-attention mechanism, where visual and language features were inputted directly into a single transformer module for information exchange.

However, due to the low computational efficiency of object detectors and the inability to update weights during multimodal pre-training, researchers have shifted towards end-to-end pre-trained models. CLIP-ViL [50] and PixelBERT [24] fed grid features extracted by CNNs and text

features into a single transformer module. ViLT [26] concatenated image patch embeddings and text token embeddings for pre-training. Recent works focus on unifying the structure of image and text encoders. For instance, ALBEF [31], METER [17], Florence [64], CoCa [63], Flamingo [2], and PaLI [10] utilized ViT [15] as an image encoder, thereby unifying the structure of image and text encoders to some extent. In contrast, VLMO [6] and BEiT-3 [58] adopted multi-way transformers to unify the modeling of text and images where text, images, and image-text pairs are fed into a single transformer module.

Knowledge Distillation. Knowledge distillation [22] is firstly proposed to transfer the knowledge learned by the teacher model to the student model. It has wide-ranging applications across various modalities [46, 25, 54, 37, 31, 59, 13] to reduce the number of parameters and improve the performance of the student model. Among these works, KD-VLP[37] proposed an object-aware end-to-end VLP framework with object knowledge distillation. CLIP-TD [59] used CLIP-targeted distillation to distill knowledge from both CLIP’s vision and language branches into the existing VL model. VLKD [13] aligned the CLIP text encoder and BART encoder to enable the capability for multimodal generation. The work is similar to our method in the multimodal field is ALBEF [31], which employed momentum distillation by using its own momentum model as the teacher model. However, while these works have focused on knowledge transfer within the same modality, our proposed *cross-distillation* method aims to transfer knowledge from a language model to a multimodal model.

Negative Sampling. Hard negative mining is a widely adopted technique to enhance model performance. Prior research [8] has demonstrated that hard negative samples have the most significant impact on the training process, rendering the easiest 95% of negatives redundant. [49, 42] generated a substantial number of hard negative samples through adversarial word replacement, resulting in a significant decline in the performance of several multi-modal models. On the other hand, works such as [19, 68, 61] employed hard negative mining to improve model performance in contrastive learning. For ITM tasks, recent works [31, 6, 16] employed hard negative mining by selecting negative samples with the highest cosine similarity to the positive samples in the same batch. Nevertheless, this method has two limitations: First, the selection of negative samples is significantly influenced by the learning of the ITC task and the batch size. Second, this method has a high probability of choosing false negative samples, thereby adversely affecting the model’s learning. To the best of our knowledge, our proposed method is the first to employ self-generated synthetic hard negative samples to enhance performance in image-text matching tasks.

3. Method

To improve vision-language pre-training, we propose ViLTA that comprises of two components: 1) Knowledge Distillation for MLM; 2) Synthetic Hard Negatives for ITM. Figure 1 shows the overall model architecture of ViLTA.

The goal of ViLTA is to further improve downstream performance of vision-language models by leveraging textual augmentation. To achieve it, the first is to utilize the frozen language encoder to generate soft labels for MLM. The second is to provide hard negatives for ITM by synthesizing negatives based on the current language encoder, which is significantly different from previous negatives selection method [31, 6].

3.1. Model Architecture

As shown in Figure 1, the overall model architecture comprises of three components, including vision encoder, language encoder, and multimodal encoder. Here, we introduce each component in detail.

Vision Encoder. We employ ViT [15] as a vision encoder to model an input image, which directly feeds image patches segmented from a whole image input and encodes them as encodes them as a sequence of embeddings $\{v_{cls}, v_1, \dots, v_N\}$ with a additional [CLS] token embedding. Following the success in previous works [50, 13, 17, 29], we initialize the weights of the image encoder using a pre-trained CLIP-ViT-224/16 model [44].

Language Encoder. We leverage RoBERTa [36] as a language encoder to model language inputs. It converts the input caption into a sequence vector $\{w_{cls}, w_1, \dots, w_N\}$, in which the embedding of the [CLS] token summarizes the global text feature. This sequence vector is then fed into the subsequent multimodal encoder to explore the relationship between image and text pairs.

Multimodal Encoder. To further capture the relationship among image-text pairs, we adopt a multimodal encoder which employs two independent cross-attention transformer modules to deeply fuse image and text information. The cross-modality multi-head attention module uses the representations of one modality as the query and another modality’s representations as the key and value, as shown in Figure 1. This deep fusion mechanism independently encodes image and text features and fuses cross-model interaction, leading to better performance improvements.

3.2. Knowledge Distillation for MLM

Masked Language Modeling (MLM) aims to predict masked words by leveraging the learned image and text features. In specific, for any certain image-text pair, we first randomly mask a portion of tokens in a sentence by substituting them with the special token [MASK]. Then, the

original masked tokens can be predicted by the remaining text input \tilde{T}_{msk} and its corresponding image input I . Thus, the MLM task can be formulated with a cross-entropy loss:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(I, \tilde{T}_{\text{msk}}) \sim D} H(y^{\text{msk}}, p^{\text{msk}}(I, \tilde{T}_{\text{msk}})) \quad (1)$$

where D represents the whole training image-text pairs, $p^{\text{msk}}(I, \tilde{T}_{\text{msk}})$ denotes the predicted probability of the masked token, and y^{msk} is a one-hot representation of the randomly masked ground-truth token.

Different from the traditional single-modality text encoders [28, 36] and previous vision-language encoders [31, 34] that randomly mask 15% of the input tokens, we increase the mask ratio from 15% to 50% in order to encourage the model to reconstruct the masked token by leveraging on both the context of text and image features. Such a mask ratio in BEiT-3 [58] also verifies that a higher mask ratio can urge the model to recover the masked token from the content of the image rather than depending only on the context of the text itself.

In the MLM task, if the model is trained to only learn one-hot labels and treat all other potential positive examples as negative examples, it could harm the model’s ability to learn effectively. In this work, we propose a novel *cross-distillation* method to generate soft labels that can replace one-hot labels. Specifically, we duplicate the language encoder as a teacher model and freeze its parameters. Next, we input the *non-masked* language sequence into the frozen language encoder to obtain a predicted probability matrix of a sequence of tokens. This is achieved by adding a masked language modeling (MLM) head on top of the output embeddings. We denote the predicted probability matrix as $q(T) \in \mathbb{R}^{S \times V}$, where S is the length of the text sequence and V is the vocabulary size. Subsequently, we select the predicted probability vector $q^{\text{msk}}(T)$ of the original masked tokens from $q(T)$ and use KL-divergence to measure the difference between the prediction of the multimodal encoder $p^{\text{msk}}(I, \tilde{T}_{\text{msk}})$ and the soft labels $q^{\text{msk}}(T)$. The distillation loss for MLM is defined as follows:

$$\mathcal{L}_{\text{dis}} = \mathbb{E}_{(I, \tilde{T}_{\text{msk}}) \sim D, T \sim D} KL(q^{\text{msk}}(T), p^{\text{msk}}(I, \tilde{T}_{\text{msk}})) \quad (2)$$

As a result, the final loss can be formulated by the combination of the original MLM loss and distill loss:

$$\mathcal{L}_{\text{mlm}}^{\text{dis}} = \alpha \mathcal{L}_{\text{mlm}} + (1 - \alpha) \mathcal{L}_{\text{dis}} \quad (3)$$

where α is a hyperparameter that controls the distillation weight. In our experiments, we set α as 0.5 for simplicity.

Discussion on Cross-Distillation. The motivation behind *cross-distillation* is rooted in the observation that when a complete caption without [MASK] token is fed into a language encoder, the encoder captures not only the contextual information surrounding each word but also information about the word itself. As shown in Figure 2, it allows

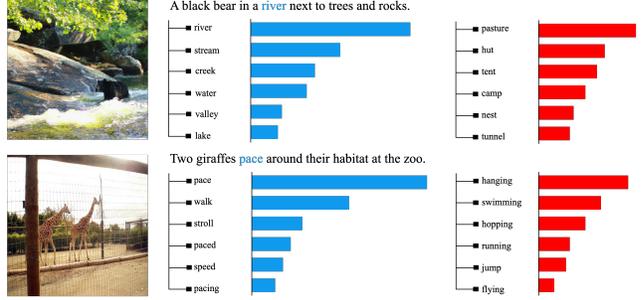


Figure 2. Examples of ViLTA are presented in two columns: 1st col displays the soft labels generated by cross-distillation, while 2nd col lists the top negative words.

the predicted tokens with high probability at each position to serve as potential synonyms or hypernyms of the original word, and the probability can be measured to determine the degree of similarity between the predicted words and the original word in context. This intuition ensures that substituting the masked token with the predicted token with a high probability will not alter the semantic meaning or grammatical structure of the sentence. Therefore, *cross-distillation* is utilized to enhance the learning efficiency, representation, and generalization ability of the model. In this approach, we combine the one-hot labels of randomly masked ground-truth tokens with the soft labels generated by a frozen language encoder to train vision-language models. This enables smooth and efficient learning, where the potential synonyms generated by the frozen language encoder can be regarded as a variant of positive samples.

3.3. Synthetic Hard Negatives for ITM

The purpose of Image-Text Matching is to capture the fine-grained alignment among image-text pairs. ITM can be viewed as a binary classification problem, which aims to predict whether an image-text pair is positive (matched) or negative (unmatched) based on the learned embeddings of the [CLS] token. Such embeddings are generated by the multimodal encoder, manifesting global cross-modality representations. Since the multimodal encoder adopts two cross-attention transformer modules, we concatenate two embeddings of the [CLS] token generated by vision and language modules respectively to obtain the final embeddings for ITM training. After that, a fully-connected (FC) layer with softmax activation function serves as a classifier to predict a two-class probability p^{itm} . The ITM loss can be represented as:

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I, T) \sim D} H(y^{\text{itm}}, p^{\text{itm}}(I, T)) \quad (4)$$

where y^{itm} denotes a binary ground-truth label.

A crucial method to improve performance on ITM is to find more informative negatives for model training. Negatives for ITM should be satisfied both of the following criteria:

- *The embedding of the hard negative sample should be similar to that of the positive sample in the embedding space.*
- *The hard negative sample must be a true negative sample with some fine-grained features that contradict the positive sample.*

To accomplish the abovementioned criteria, we propose to synthesize hard negatives through textual augmentation, which differs significantly from ALBEF [31] selecting negatives with higher contrastive similarities from ITC in the current batch. Such a synthetic method can generate negatives that are close to the positive sample in the embedding space while alleviating simultaneously the false negative issue. Specifically, for any given positive image-text pair (I, T) , the synthetic hard negatives method generates its corresponding negative pair (I, T_n) , which can be divided into three steps:

1) Generate the masked sentence T' : Substituting one word with a [MASK] tag in the original sentence T , based on part-of-speech (POS) tagging. The POS tagging helps identify crucial parts of a sentence, such as nouns, verbs, adjectives, and numerals.

2) Calculate predicted probability p_m : Taking the masked sentence T' as input and passing it into the current language encoder to compute probabilities $p_m(t|T')$ at the [MASK] position during the beginning of each training step.

3) Synthesize hard negative sentence T_n : Sampling the probability of the corresponding [MASK] position $q_m^{\text{msk}}(T)$ from $q(T)$ which we have used in cross distillation module; calculating the negative sampling distribution $p_n = \frac{p_m(t|T')}{q_m^{\text{msk}}(T)}$; sampling one word based on p_n to synthesize hard negative sentence T_n .

Discussion on Hard Negatives. In contrast to prior works that sample hard negatives from existing negatives, our proposed ViLTA paradigm synthesizes negatives based on the current language encoder. As mentioned in the previous section, the plausibility of predicted words in the masked position can be measured by $p_m(t|T')$, while $q_m^{\text{msk}}(T)$ can gauge the similarity between the prediction words and the original word (See Figure 2). Consequently, selecting the predicted word with the highest p_n ensures its plausibility in the context and its dissimilarity with the original word. This approach introduces a novel perspective on hard negatives, as it can be seen as a variant of data augmentation. Specifically, instead of generating positive samples for contrastive learning [20, 9] through data augmentation, we aim to synthesize hard negative samples for the ITM objective in vision-language pretraining. These synthetic hard negatives offer additional information for model training since they conform to, but differ from, the positive samples, thereby

accelerating the model’s convergence and enhancing downstream performance.

4. Experiments

To demonstrate the effectiveness of ViLTA, we conduct comprehensive experiments on 5 vision-language tasks. First, we introduce experimental setup, including model architecture, pre-training datasets, downstream tasks, and implementation details. Second, we compare our proposed ViLTA with other classical vision-language pre-training models on various tasks, including visual question answering, visual reasoning, image-text retrieval, image captioning. Lastly, we design a series of ablation studies for model analysis.

4.1. Experimental Setup

Pre-training Datasets. In pre-training, we collect a vast number of image-text pairs from the Internet to train our model with two pre-training tasks (MLM and ITM). In line with previous research, we adopt four datasets for pre-training, including Conceptual Captions [48], COCO [35], SBU Captions [41], and Visual Genome [27]. Here, we utilize 4 million images for training since a large proportion of the image links have broken in the process of downloading datasets. Statistics of datasets are shown in Appendix A.

Downstream Tasks. We evaluate our proposed ViLTA on 5 downstream vision-language tasks, including visual question answering, visual reasoning, visual entailment, image-text retrieval, and image captioning tasks. The detailed description of each task is represented in Appendix B.

Implementation Details. We train ViLTA on 8 NVIDIA A100 GPUs for a total of 360,000 steps with a batch size of 1024, which takes a period of approximately 5 days. The maximum text sequence length is set to 50 and a max resolution of pre-training images is set to 288×288 . We utilize the AdamW optimizer [38] with an initial learning rate of $1e - 5$ for the bottom vision and language encoders and $5e - 5$ for the top multimodal encoder. To optimize the learning process, we adopt a linear decay learning rate schedule that contains a warm-up period with a ratio of 10%. The learning rate is subsequently linearly decreased to $1e - 8$ after 10% of the total training steps.

4.2. Results on VL Classification Tasks

We conduct an empirical evaluation of our proposed ViLTA on vision-language (VL) classification tasks, including VQA, visual reasoning, and visual entailment. In order to demonstrate the effectiveness of ViLTA, we compare it with SOTA methods and report the experimental results in Table 1. It can be obviously observed that ViLTA achieves impressive performances on the VQAv2 dataset and out-

Model	#Pretrain Images	Visual Encoder	VQAv2		NLVR ²		SNLI-VE	
			test-dev	test-std	dev	test	dev	test
<i>BASE-Size Models</i>								
ViLT [26]	4M	VIT-B-384/32	71.26	-	75.70	76.13	-	-
UNITER _{BASE} [11]	4M	Faster R-CNN	72.70	72.91	77.18	77.85	78.59	78.28
GLIPv2 _{BASE} [66]	20M	Swin-B-224	73.1	73.3	-	-	-	-
VILLA _{BASE} [18]	4M	Faster R-CNN	73.59	73.67	78.39	79.30	79.47	79.03
UNIMO _{BASE} [33]	4M	Faster R-CNN	73.79	74.02	-	-	80.00	79.10
CLIP-ViL _{BASE} [50]	9.2M	CLIP-Res50	73.92	74.09	-	-	78.64	78.97
KD-VLP [37]	4M	ResNet-101	74.20	74.31	77.36	77.78	78.21	77.87
ALBEF [31]	4M	DEIT-B-224/16	74.54	74.70	80.24	80.50	80.14	80.30
ALBEF [31]	14M	DEIT-B-224/16	75.84	76.04	82.55	83.14	80.80	80.91
VinVL _{BASE} [67]	5.7M	ResNeXt-152	75.95	76.12	82.05	83.08	-	-
VLM _{BASE} [6]	4M	MOME Transformer	76.64	76.89	82.77	83.34	-	-
BLIP _{BASE} [30]	14M	DEIT-B-224/16	77.54	77.62	82.67	82.30	-	-
METER-CLIP-ViT [17]	4M	CLIP-ViT-B-224/16	77.68	77.64	82.33	83.05	80.86	81.19
SimVLM _{BASE} [60]	1.8B	ResNet-101	77.87	78.14	81.72	81.77	84.20	84.15
OFA _{BASE} [57]	54M	ResNet-101	77.98	78.07	-	-	89.30 [†]	89.20 [†]
X-VLM [65]	4M	Swin-B-224	78.07	78.09	84.16	84.21	-	-
BLIP _{BASE} [30]	129M	DEIT-B-224/16	<u>78.25</u>	<u>78.32</u>	82.15	82.24	-	-
ViLTA _{BASE}	4M	CLIP-ViT-B-224/16	78.62	78.47	<u>83.21</u>	84.29	<u>81.50</u>	<u>81.67</u>
<i>Large-Size Models</i>								
UNITER _{LARGE} [11]	4M	Faster R-CNN	73.82	74.02	79.12	79.98	79.39	79.38
VILLA _{LARGE} [18]	4M	Faster R-CNN	74.69	74.87	79.76	81.47	80.18	80.02
UNIMO _{LARGE} [33]	4M	Faster R-CNN	75.06	75.27	-	-	81.11	80.63
VinVL _{LARGE} [67]	5.7M	ResNeXt-152	76.52	76.60	82.67	83.98	-	-
CLIP-ViL _{LARGE} [50]	9.2M	CLIP-Res50×4	76.48	76.70	-	-	80.61	80.20
VLM _{LARGE} [6]	4M	MOME Transformer	<u>79.94</u>	<u>79.98</u>	85.64	86.86	-	-
ViLTA _{LARGE}	4M	CLIP-ViT-L-334/14	80.19	80.17	<u>85.16</u>	<u>86.13</u>	83.12	82.98
<i>Huge-Size Models</i>								
SimVLM _{HUGE} [60]	1.8B	Larger ResNet-152	80.03	80.34	84.53	85.15	86.21	86.32
BEIT-3 [58]	28M	MOME Transformer	84.19	84.03	91.51	92.58	-	-
PaLI [10]	1.6B	VIT-E-224	84.30	84.34	-	-	-	-

Table 1. Result comparison with representative vision-language pre-training models. [†] denotes using additional text premise as input.

performs all baselines on both test-dev and test-std with either BASE or LARGE architecture. Notably, with the same amount of pre-training images (4M), ViLTA significantly surpasses other models [50, 17] adopted CLIP-weights to initialize the vision encoder. Additionally, ViLTA brings about performance improvements in visual reasoning and visual entertainment over most of baselines, especially with the condition of the same amount of pre-training images. It clearly demonstrates that the superiority of our proposed ViLTA in VL classification tasks. By performing BASE and LARGE architectures on the downstream classification tasks, the experimental results indicate that scaling the model’s parameters can result in a significant performance improvements, which provides additional experimental sup-

ports to verify the effectiveness of ViLTA.

4.3. Results on VL Retrieval Tasks

Table 13 shows the results on VL retrieval tasks. We can find that ViLTA outperforms existing VL models in most cases, especially on the Flickr dataset. Notably, the results in recent studies [31, 30, 16] illustrate the importance of ITC for retrieval tasks, which integrates ITC task into the pre-training and adopts a re-ranking strategy in the fine-tuning. Such investigations reveal that ITC can bring about a consistent improvement of 6% in terms of R@1. The above fact may lead to slight performance improvements on the Flickr dataset and even performance degradation on the COCO dataset since ViLTA only adopts MLM and ITM

Method	Flickr						COCO					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ViLT _{BASE} [26]	64.4	88.7	93.8	83.5	96.7	98.6	42.7	72.9	83.1	61.5	86.3	92.7
PixelBERT [24]	71.5	92.1	95.8	87.0	98.9	99.5	50.1	77.6	86.2	63.6	87.5	93.6
UNITER _{BASE} [11]	72.5	92.4	96.1	85.9	97.1	98.8	50.3	78.5	87.2	64.4	87.4	93.1
VILLA _{BASE} [18]	74.7	92.9	95.8	86.6	97.9	99.2	-	-	-	-	-	-
OSCAR [34]	-	-	-	-	-	-	54.0	80.8	88.5	70.0	91.1	95.5
VLMO _{BASE} [6]	79.3	95.7	97.8	92.3	99.4	99.9	57.2	82.6	89.8	74.8	93.1	96.9
ALBEF _{BASE} [31]	82.8	96.7	98.4	94.3	99.4	99.8	56.8	81.5	89.2	73.1	91.4	96.0
METER-CLIP-ViT [17]	82.2	96.3	98.3	94.3	99.6	99.9	57.1	82.7	90.0	76.2	93.2	96.8
ViLTA _{BASE}	85.2	97.2	98.8	94.5	99.8	99.8	59.5	83.1	89.7	73.3	91.8	95.9

Table 2. Experimental results on image retrieval (IR) and text retrieval (TR) on Flickr30K and COCO datasets.

Method	COCO				NoCaps Val		NoCaps Test	
	BLEU@4	METEOR	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
CLIP-ViL-ViT [50]	21.1	19.4	58.0	12.2	-	-	-	-
GLIPV2 _{BASE} [66]	37.4	-	123.0	21.9	-	-	-	-
UFO _{BASE} [56]	36.0	28.9	122.8	22.2	80.7	12.5	78.8	12.5
VinVL _{BASE} [67]	38.2	30.3	129.3	23.6	94.3*	13.1*	92.5*	13.1*
METER-CLIP-ViT [17]	38.8	30.0	128.2	23.0	-	-	-	-
FIBER [16]	39.1	30.4	128.4	23.1	88.6	13.0	86.0	12.9
SimVLM _{BASE} [60]	39.0	32.9	134.8	24.0	-	-	-	-
LEMON _{BASE} [23]	40.3	30.2	133.3	23.3	100.4	13.8	-	-
ViLTA _{BASE}	41.0	30.9	135.1	23.6	103.4	13.9	100.2	13.9

Table 3. Results on image captioning. * denotes that the model was optimized using the CIDEr metric to improve performance. It is worth noting that all the results reported for our model were obtained without CIDEr optimization.

tasks for pre-training.

4.4. Results on Image Captioning

The experimental results in Table 3 demonstrate that ViLTA achieves better performance in the image captioning task. In specific, we fine-tune the model on the COCO Captions [35] dataset and evaluate its performance with four metrics such as BLEU@4, METEOR, CIDEr, and SPICE. From Table 3, we can observe that ViLTA consistently outperforms all baselines, especially compared to CLIP-ViL. Such results verify the effectiveness of ViLTA rather than the importance of the initialization of CLIP-ViT. Furthermore, we conduct experiments on the NoCaps [1] dataset without any additional fine-tuning or optimization techniques. It indicates that ViLTA can urge the vision-language model to recognize fine-grained objects to enhance caption quality.

4.5. Model Analysis

To further analyze the impact of each component of ViLTA, we perform a series of experiments with various objectives on VL classification and image captioning tasks.

Table 4 demonstrate the performance comparison among various variants of our proposed ViLTA. Compared to the basic variant (MLM), incorporating the ITM task can benefit all downstream vision-language tasks. Such benefits can be further enhanced with synthetic hard negatives and *cross-distillation* method respectively. Synthetic hard negatives are utilized to boost the ITM tasks while *cross-distillation* method is leveraged to improve the learning of the MLM task. We can find that the performance improvements of hard negatives are higher than *cross-distillation*. By leveraging these two vital techniques, ViLTA can achieve the best overall performance among different variants.

Objectives	VQAv2	NLVR2	SNLI-VE	COCO _{Cap}
MLM	76.65	82.21	80.16	38.2
MLM+ITM	77.34	82.97	80.90	38.8
MLM+ITM _{hard}	78.06	83.80	81.48	39.8
MLM _{distill} +ITM	77.85	83.52	81.22	39.6
ViLTA _{BASE}	78.47	84.29	81.67	40.4

Table 4. Impact of each component in ViLTA.

4.6. Ablation Studies

Analysis on Image Captioning. We conduct an ablation study that aims to enhance the adaptability of pre-trained models for image captioning tasks in Table 5. Specifically, we introduce an additional pre-training phase that involved a language modeling (LM) task on the 4 million dataset for one epoch. Our results show that a improvement of 0.6 in BLEU@4 when incorporating the LM task. Furthermore, we investigate the issue of information leakage caused by deep fusion in the cross-attention modules. We explore two approaches to address this problem: removing the cross-attention module on top of the image encoder or transforming it into a self-attention module. We compare their performance and conclude that the cross-attention module should be retained for optimal performance.

COCO	w/ LM		w/o LM	
	w/o CA	w/ CA	w/o CA	w/ CA
BLEU@4	40.0	41.0	39.6	40.4
CIDEr	131.5	135.1	130.0	133.1

Table 5. Ablation study on image captioning. LM: language modeling pre-training. CA: cross-attention module.

Impact of Negative Mining Method. We conduct an experiment to analyze the impact of different negative mining methods. One straightforward approach is to utilize WordNet [40] to generate negative samples by substituting antonyms from the original sentence. However, the experimental results in Table 6 demonstrate that WordNet brings about a minimal performance improvement compared with in-batch randomly selected negative samples. Besides, pre-generated negatives are preprocessed at the beginning of pre-training, which performs better than WordNet but worse than dynamically-generated negatives. The possible reason is that dynamically-generated negatives are updated at each step based on the language encoder, ensuring the hardness of negative samples. More importantly, the combination of WordNet and dynamically-generated negatives shows performance degradation compared to the raw dynamically-generated negatives. Such finding can be attributed to the fact that the negatives generated by WordNet differ from the original sentence in the semantic space and representation space, making it easier for the model to distinguish them.

4.7. Efficiency Analysis

ViLTA employs two distinct strategies to enhance downstream performance: the *cross-distillation* technique in the Masked Language Modeling (MLM) task and the synthetic hard negative mining technique in the Information-Theoretic Metric (ITM) task. To shed light on the training cost associated with these techniques, we present the

Negative Mining Method	VQAv2	
	test-dev	test-std
In-batch random	74.08	74.37
WordNet	74.24	74.36
Pre-generated	74.52	74.72
WordNet+Dynamically-generated	74.71	74.86
Dynamically-generated	74.93	75.12

Table 6. Ablation study on negative mining method. In the pre-generated approach, all negative samples are generated before the training phase, while in the dynamically-generated approach, negative samples for each batch are updated during each training step.

Pre-training Strategies		GPU-hours
cross-distill	hard negative	A100
✗	✗	34.68
✓	✗	37.56
✗	✓	38.09
✓	✓	38.43

Table 7. The impact of different pre-training strategies on the time consumption of each epoch evaluated with a single A100 GPU.

findings in Table 7. The combined implementation of both strategies introduces a marginal increase of approximately 10% in training time per epoch. It is noteworthy, however, that this increment in training time is inconsequential in light of the benefits accrued. The adoption of these two techniques not only offsets the relatively minor increase in training duration but also contributes to the acceleration of model convergence.

4.8. Case Study

To demonstrate the effectiveness of ViLTA in accurately identifying objects, attributes, actions, and quantitative relationships, we utilize Grad-CAM [47] for visualizations. Figure 3 presents the results for the VQA task, indicating that our model can focus on fine-grained features in the image and correctly answer questions based on them. Besides, as shown in Figure 4, visualizations for the retrieval task on a per-word basis illustrate the model’s precise identification of objects and actions within the image, illustrating the model’s precise identification of objects and actions within the image.

5. Conclusion

In this paper, we propose a novel vision-language pre-training method ViLTA to further improve the representability of the model. Specifically, we propose a *cross-distillation* method to generate soft labels to address the issue of treating synonyms of the masked words as negative samples in one-hot labels for MLM, which improves the

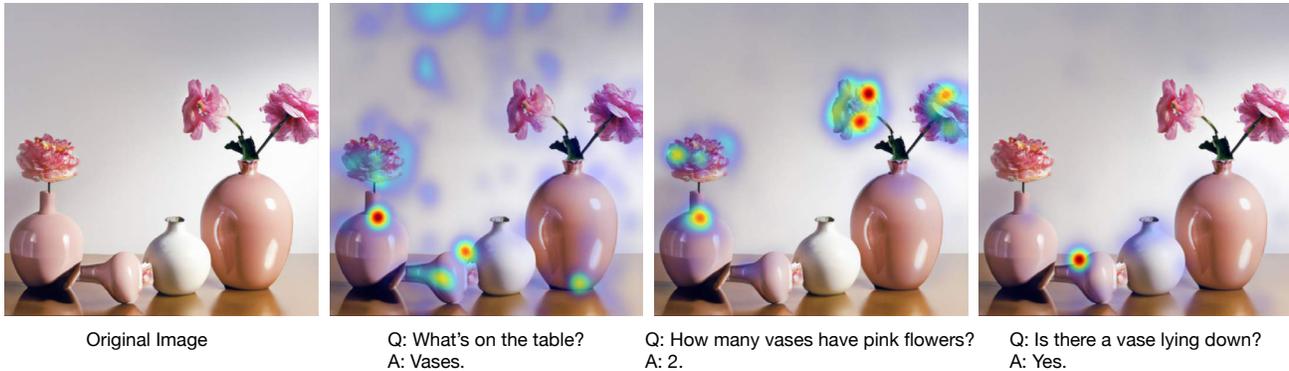


Figure 3. The Grad-CAM visualization of VQA.



Figure 4. The Grad-CAM visualization of words in the caption “A white bird perched on a rock by the ocean.”

robustness of vision-language models. Moreover, we design a new negative selection method for ITM, which aims to synthesize hard negatives based on the current language encoder by leveraging the context of language input. Such hard negatives provide more information for model convergence, which significantly enhances the downstream performances. Extensive experimental results on five vision-language tasks demonstrate the effectiveness and generalization ability of the proposed method.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62277033). It also got partial support from National Engineering Laboratory for Cyber-learning and Intelligent Technology, and Beijing Key Lab of Networked Multimedia.

References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [6] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.
- [7] Emanuele Bugliarelli, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. Measuring progress in fine-grained vision-and-language understanding. *arXiv preprint arXiv:2305.07558*, 2023.
- [8] Tiffany Tianhui Cai, Jonathan Frankle, David J Schwab, and Ari S Morcos. Are all negatives created equal in contrastive instance discrimination? *arXiv preprint arXiv:2010.06682*, 2020.

- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [10] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.
- [12] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.
- [13] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [16] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022.
- [17] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.
- [18] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.
- [19] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pages 752–768. Springer, 2020.
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [21] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.
- [22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [23] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022.
- [24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.
- [25] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.
- [26] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.
- [27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- [28] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- [29] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.
- [30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [31] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [32] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [33] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal

- contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.
- [34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.
- [35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [37] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [40] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [41] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [42] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.
- [43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [46] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypemymed, 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)*, pages 2556–2565, 2018.
- [49] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.
- [50] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.
- [51] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [52] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Hua-jun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.
- [53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [56] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.
- [57] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [58] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a

- foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.
- [59] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022.
- [60] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.
- [61] Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, and Stan Z Li. Progccl: Rethinking hard negative mining in graph contrastive learning. In *International Conference on Machine Learning*, pages 24332–24346. PMLR, 2022.
- [62] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.
- [63] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [64] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- [65] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.
- [66] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022.
- [67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.
- [68] Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. *arXiv preprint arXiv:2104.06245*, 2021.

Appendix

A. Pre-training Details

The statistics of pre-training datasets are presented in Table 8. The COCO Captions dataset comprises manually generated captions where multiple captions are assigned to each image. For the Visual Genome dataset, the region description serves as the image caption, yielding several captions for each image. The SBU Captions and Conceptual Caption datasets contain a single caption per image. It should be noted that a considerable number of the image links in these two datasets have become invalid because they are collected from the Internet.

	COCO	VG	SBU	CC3M
#Images	113K	108K	855K	2.98M
#Captions	567K	5.4M	855K	2.98M

Table 8. Statistics of the pre-training datasets.

The default architecture of ViLTA contains a dual-encoder architecture (a pre-trained vision encoder and a pre-trained language encoder) and a multimodal encoder. Table 9 reports the hyperparameters used in our pre-training model. For ViLTA_{BASE}, we leverage a 12-layer transformer-based structure as language/vision encoder and 6-layer for the multimodal encoder respectively. The number of transformer layers for the language and vision encoders is set to 24 for ViLTA_{LARGE}. The number of the multimodal encoder also maintains the default setup of 6-layer transformer-based structure. Here, we initialize the language encoder with weights from the pre-trained RoBERTa [36] and the vision encoder with weights from the pre-trained CLIP-ViT-224/16 [44].

B. Fine-tuning Details

We fine-tune ViLTA on 5 downstream tasks using the hyperparameters reported in Table 10 for VL classification tasks, Table 11 for VL retrieval tasks, Table 12 for image captioning. In the following sections, we provide a comprehensive description of the fine-tuning configurations employed for each task.

- *Visual Question Answering (VQA)* [4] aims to predict a natural language answer based on the given image and question. Following the previous works [26, 17, 6, 58], we treat VQA as a multi-label classification task with 3,129 possible answers. We concatenate the image representation v_{cls} and text representation w_{cls} obtained from the multimodal model, and then pass it through a 2-layer MLP layer to perform a classification task. We use GELU activation function and a binary

Hyperparameters	ViLTA _{BASE}	ViLTA _{LARGE}
Total steps	36k	24k
Warmup steps	21.6k	14.4k
Batch size	1024	1024
Learning rate	$1e^{-5}$	$4e^{-6}$
Learning rate decay		Linear
Weight decay		0.01
Dropout ratio		0.1
AdamW ϵ		$1e^{-8}$
AdamW β		(0.9, 0.98)
Textual encoder	RoBERTa _{BASE}	RoBERTa _{LARGE}
Visual encoder	CLIP-ViT-B-224	CLIP-ViT-L-336
Patch size	16	14
Input resolution	288	224
Number of layers	6	6
Hidden size	768	1024
FFN inner hidden size	3072	4096
Number of attention heads	12	16

Table 9. Hyperparameters for pre-training model. The last block is the hyperparameters for the multimodal encoder.

cross-entropy loss function on the soft target scores to optimize the model.

- *Visual Reasoning* focuses on predicting whether the caption is true or false for a pair of images. Here, we employ a pairwise strategy to effectively process the input in NLVR² [52] dataset, where each data sample is divided into (*image1*, *statement*) and (*image2*, *statement*). We then feed them separately into the model to obtain two representations and concatenate them together to pass through a binary classification head.
- *Visual Entailment* aims to predict whether a natural language hypothesis is entailed, neutral or contradicted by the image premise. We train and evaluate our model on SNLI-VE [62] dataset and treat it as a three-class classification problem.
- *Image-Text Retrieval* contains two sub tasks: image-to-text retrieval (TR) and text-to-image retrieval (IR). COCO [35] and Flickr30K [43] serve as evaluation datasets. Following the standard setting in ViLT [26], We use the pre-trained ITM head, specifically the component that calculates the true-pair logits, to initialize the similarity score head. We then sample 15 random texts as negative examples and use a cross-entropy loss that maximizes the scores for positive pairs.
- *Image Captioning* is a generative task and we inves-

tigate whether our encoder-only model is suitable for such generative tasks. To adapt our model for image captioning, we modify the encoder on the text side of the model by transforming it into a causal decoder through the adjustment of the attention mask. Subsequently, we fine-tune the model on the COCO Captions [35] dataset using cross-entropy loss and evaluate it on the NoCaps [1] dataset without additional training.

Hyperparameters	VQAv2	NLVR ²	SNLI-VE
Learning rate	$1e^{-5}$	$1e^{-5}$	$2e^{-6}$
Epochs	10	10	5
Batch size	512	256	64
AdamW ϵ		$1e^{-8}$	
AdamW β		(0.9, 0.98)	
Weight decay	0.05	0.01	0.01
Dropout ratio		0.1	
Input resolution	576^2	384^2	288^2

Table 10. Hyperparameters for fine-tuning ViLTA on VL classification tasks.

Hyperparameters	COCO Flickr
Learning rate	$5e^{-6}$
Epochs	10
Batch size	64
AdamW ϵ	$1e^{-8}$
AdamW β	(0.9, 0.98)
Weight decay	0.01
Dropout ratio	0.1
Input resolution	576^2

Table 11. Hyperparameters for fine-tuning ViLTA on VL retrieval tasks.

C. Scaling Ability

To show the effectiveness of ViLTA on extensive datasets, we expand the training of ViLTA-base and ViLTA-large on a subset of the LAION-2B and CC12M datasets employing 64 A100 GPUs in Table 13. The total volume of data was roughly 150M, comparable to the 129M dataset used in BLIP. All performance metrics for retrieval tasks show substantial enhancements, ranging from 73.3 to 80.5 on the COCO dataset for text retrieval in terms of recall@1. However, the gain in VL understanding (VLU) tasks is not as prominent as the increase in retrieval tasks, which is consistent with the findings in previous studies [30, 7]. Such

Hyperparameters	COCO Captioning
Learning rate	$1e^{-5}$
Epochs	10
Batch size	512
AdamW ϵ	$1e^{-8}$
AdamW β	(0.9, 0.98)
Weight decay	0.01
Dropout ratio	0.1
Input resolution	576^2
Label smoothing ϵ	0.1
Beam size	5

Table 12. Hyperparameters for fine-tuning ViLTA on image captioning.

discrepancy arises due to the challenges associated with the considerable noise present in large-scale web data, which are integral to VLU tasks. As shown in Table 14, in the context of a large-scale dataset, ViLTA achieves a better gain, while, in contrast, BLIP brings about performance degradation.

Dataset	Flickr			COCO		
	TR@1	TR@5	TR@10	TR@1	TR@5	TR@10
4M	94.5	99.8	99.8	73.3	91.8	95.9
150M	95.7	99.9	99.9	80.5	94.6	97.3

Table 13. Experimental results on retrieval task.

Dataset	BLIP		ViLTA	
	14M	129M	4M	129M
NLVR2-dev	82.67	82.15	85.16	86.33
NLVR2-test	82.30	82.24	86.13	87.25

Table 14. Results on NLVR2 dataset. Large scale data may not have significant benefits for VLU tasks.

D. Additional Results

In this section, we present additional results generated by ViLTA. Specifically, we show the efficacy of ViLTA in image captioning. The case study in Figure 5 shows the generated image captions on a series of samples. Notably, ViLTA generates diverse and descriptive captions, which can effectively encapsulate the content of the corresponding images. These results verify the effectiveness of ViLTA in different VL tasks.



A white train traveling down a street next to a tall clock tower.



A white and black fire hydrant in a parking lot.



A row of surfboards sticking out of the sand.



A man flying through the air while riding a skateboard.



A bunch of umbrellas that are hanging from the ceiling.



A sandwich cut in half on a plate.



A herd of sheep standing on top of a lush green field.



A teddy bear sitting on top of a pole.



A street scene with cars and traffic lights.



A young boy holding a Nintendo Wii game controller.



A man riding a dirt bike on top of a lush green field.



Three giraffes are standing in a grassy field.

Figure 5. Case study of ViLTA on image captioning task.