LightVLP: A Lightweight Vision-Language Pre-training via Gated Interactive Masked AutoEncoders

Xingwu Sun[1,2], **Zhen Yang**[1], **Ruobing Xie**[1*], **Fengzong Lian**[1], **Zhanhui Kang**[1], **Chengzhong Xu**[2]
[1] Machine learning platform department, Tencent
[2] University of Macau
{sammsun,ruobingxie}@tencent.com

Abstract

This paper studies vision-language (V&L) pre-training for deep cross-modal representations. Recently, pre-trained V&L models have shown great success in V&L tasks. However, most existing models apply multi-modal encoders to encode the image and text, at the cost of high training complexity because of the input sequence length. In addition, they suffer from noisy training corpora caused by V&L mismatching. In this work, we propose a lightweight vision-language pre-training (LightVLP) for efficient and effective V&L pre-training. First, we design a new V&L framework with two autoencoders. Each autoencoder involves an encoder, which only takes in unmasked tokens (removes masked ones), as well as a lightweight decoder that reconstructs the masked tokens. Besides, we mask and remove large portions of input tokens to accelerate the training. Moreover, we propose a gated interaction mechanism to cope with noise in aligned image-text pairs. As for a matched image-text pair, the model tends to apply cross-modal representations for reconstructions. By contrast, for an unmatched pair, the model conducts reconstructions mainly using uni-modal representations. Benefiting from the above-mentioned designs, our base model shows competitive results compared to ALBEF while saving 44% FLOPs. Further, we compare our large model with ALBEF under the setting of similar FLOPs on six datasets and show the superiority of LightVLP. In particular, our model achieves 2.2% R@1 gains on COCO Text Retrieval and 1.1% on refCOCO+.

Keywords: Vision-language pre-training, Lightweight V&L pre-training, Mask autoencoder

1. Introduction

Self-supervised learning is becoming a dominating paradigm in computer vision and natural language processing. It leverages large-scale unlabeled corpora and learns knowledge through self-supervised tasks. Recently, vision-language (V&L) pre-training has attracted more attention with the availability of more image-text parallel datasets, e.g., Lin et al. (2014), Sharma et al. (2018), and Krishna et al. (2017). It learns cross-modal interactions between image and natural language using well-designed tasks such as masked language modeling, masked image modeling and image-text matching to improve downstream V&L tasks.

There exist four main challenges in V&L pre-training: (1) The training process is computationally expensive. As we know, self-attention in Transformer is quadratic time complexity to the sequence length while the input sequences of V&L models are usually long by taking in both image and text tokens. (2) The self-supervised tasks are keys to obtaining good understanding ability. How to design more challenging tasks needs further studying. (3) The embeddings from two modalities tend to reside in their own space, making it hard for information fusion. (4) Most training corpora (image-text pairs) are web noisy data, and thus it is essential to cope with the unrelated image-text pairs. Uni-modal representations should be more highlighted with mismatched image-text examples.

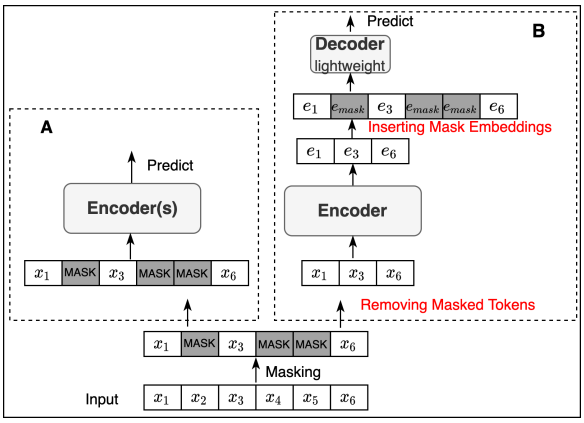

Figure 1: An example of our mask-and-remove technique. Part A shows the existing pre-training architecture and Part B is our LightVLP. Removing masked tokens will reduce the input length of VLP and thus improve the efficiency.

The development of V&L pre-training (VLP) models can be roughly divided into three stages, i.e., single-stream encoders, dual encoders, and hybrid methods. (1) Models like Qi et al. (2020) and Su et al. (2019) study on single-stream encoders to solve the third challenge by taking in the con-

[*] Ruobing Xie is the corresponding author.

LREC-COLING 2024, pages 10499–10510
20-25 May, 2024.

catenated image and text tokens and computing cross-modal interactions. Though effective, one main limitation of these models lies in the high training cost. (2) Radford et al. (2021) and Jia et al. (2021) propose dual encoders to resolve the first challenge for more efficient models. In addition, they introduced contrastive learning, which is an effective task especially for retrieval tasks. However, classical dual encoders fail to model deep interactive signals between two modalities. (3) Tan and Bansal (2019) and Li et al. (2021) propose hybrid methods to utilize the advantages of cross-modal encoders (for more comprehensive V&L interaction modeling) and dual encoders (for efficiency). In addition, Li et al. (2021) presents a distillation method to tackle the fourth challenge by computing soft pseudo targets for each pre-training task. These VLP models were supreme before the era of the dominating usage of large pre-trained models.

Despite the success of the hybrid method, it can be further improved regarding the above challenges. (1) As in Fig. 1 Part A, the model takes the full-length sequence as input without removing the masked ones. Hence, the computation complexity is still high. (2) It needs further explorations to find suitable masking rates for the masked image modeling and masked language modeling tasks, balancing both effectiveness and efficiency. (3) Existing models are still struggling with intrinsic noises brought by the unrelated/mismatched image-text pairs in the training corpora. They still simply fuse the image-text representations to predict masked tokens without considerations of confidence, inducing noise for V&L modeling.

To overcome the above-mentioned limitations, we propose a simple but effective autoencoder based V&L pre-training model, named **LightVLP**. (1) LightVLP has two interactive autoencoders for visual and textual modeling, each comprising a lightweight encoder and a decoder as shown in Fig.1 Part B. In the encoder, by masking and removing a large ratio of image and text tokens, we reduce the input sequence length and computational cost by a large margin. (2) We explore different masking ratios in the proposed framework to balance computational cost and model performance. (3) We present a gated interaction mechanism to deal with noise in image-text pairs, in which a gate is designed to choose between cross-modal representations and uni-modal representations. Intuitively, for a matched image-text pair, the model tends to choose cross-modal representations for reconstructions in both autoencoders. On the contrary, for an unmatched image-text pair, the model is more likely to predict based on the uni-modal representations without cross-modal assistance. We highlight the lightness of our proposed LightVLP compared to recent powerful large pre-trained model enhanced VLP. Compared to models like BLIP-2 (Li et al., 2023) that inserts image representations into the input of LLMs, LightVLP is more suitable for retrieval tasks, since the modality embeddings could be pre-calculated and fast retrieved in online retrieval. Compared to models such as BEIT-3 (Wang et al., 2022) and VLMO (Bao et al., 2022) that rely on certain customized decoupled modality-aware experts for different tasks, LightVLP is more flexible and lightweight, which could be flexibly adopted with other (lighter or heavier) encoder architectures and single-modality/aligned training data. In the era of LLMs, LightVLP is also valuable for the usage with scarce resources and flexible demands.

In implementation, we pre-train two types of models, LightVLP$_{base}$ and LightVLP$_{large}$. Compared to ALBEF (Li et al., 2021) with similar parameters, the LightVLP$_{base}$ model saves 44% computational cost by FLOPs while performing competitively (masking 50% image and 50% text tokens). Besides, we train a LightVLP$_{large}$ model of similar FLOPs as ALBEF and it shows significant improvements on five V&L benchmarks. In addition, we experiment on different masking settings 25%, 50% and 75%, and give the most recommended masking rates.

Our contributions can be summarized as follows.

- We propose LightVLP, a new framework for V&L pre-training. In LightVLP, we mask and remove a large portion of the input sequence to reduce computational complexity.
- We propose a gated interaction mechanism to tackle noise in unmatched image-text pairs. As for a matched image-text pair, the model computes cross-modal embeddings for reconstructions. Otherwise, the model should more rely on uni-modal embeddings.
- We explore different settings of masking ratios for different modalities by thorough experiments to reach the balance between computational cost and performance, which will inspire similarly structured VLP applications.
- We conduct experiments on six commonly used benchmarks. The results display significant improvements compared to the competitive and similar-scale model ALBEF, e.g., 2.2% improvements on COCO Text Retrieval by R@1 and 1.1% on refCOCO+.

2. Related Work

In recent years, model pre-training has attracted increasing interests in the field of natural language processing (e.g., Devlin et al. (2018), Liu et al. (2019b)), computer vision (e.g., Bao et al. (2021), Touvron et al. (2021), Dosovitskiy et al. (2020)) as

well as V&L understanding (e.g., Gan et al. (2020), Tan and Bansal (2019), Li et al. (2020b) and Lu et al. (2020)). Existing V&L models can be roughly divided into three stages, single-stream encoders, dual encoders and hybrid methods.

Single-stream Encoders concatenate the text tokens and the image features (either from object detector or image pixels) and apply a Transformer-based encoder to model their deep interactions. Su et al. (2019) proposed VL-BERT and Li et al. (2019) raised VisualBERT. These models took the text and region-of-interest of images as input and then applied a bi-encoder to encode the concatenated image-text sequences. Qi et al. (2020) presented ImageBERT, which employed an object detector to get objects from images for pre-training. Unicoder-vl (Li et al., 2020a) and UNITER (Chen et al., 2020) were models that applied the language model architecture for pre-training and could be used for cross-modal generation tasks. Yu et al. (2021) proposed to enhance knowledge into cross-modal pre-training, named ERNIE-Vil. Tan and Bansal (2019) also proposes LXMERT. Li et al. (2020c) gave OSCAR by an object and word alignment method. Kim et al. (2021) proposed to delete object detection in the V&L pre-training process. Though effective, this kind of model may fail to distinguish between intra-modality interactions and cross-modal interactions. In addition, the model is inefficient and unapplicable in real applications especially for the Image Retrieval task.

Dual Encoders apply two separate encoders to compute the image and text representations. Models equipped with dual encoders usually apply shallow interactions or dot production to project two modalities into one common semantic space, which are more efficient, while they lack the ability to model deep interactions between two modalities. As a result, their performance is no better than the above-mentioned single-stream models. Radford et al. (2021) and Jia et al. (2021) proposed to improve V&L pre-training by incorporating contrastive learning. They introduced negative samples and designed a contrastive task for better cross-modal understanding. This method is more efficient, but lacks deep cross-modal interactions.

Hybrid Methods are proposed to leverage the advantages of the two above-mentioned methods. The hybrid methods apply separate encoders to get individual representations and then employ deep cross attention layers to conduct interactions. Lu et al. (2019) first applied a two-stream structure to extract features and then used cross Transformer for information fusion. Tan and Bansal (2019) proposed a framework with three encoders to get the representations for objects, texts and their connections. Li et al. (2021) presented the ALBEF model, which aligns the vision and language tokens before interactive computation. Singh et al. (2022) gave a foundational V&L model that can learn good vision, language and cross representations at the same time. Bao et al. (2022) showed a method based on mixture-of-experts to jointly learn dual encoders and a cross encoder, which can be used separately in downstream tasks. BEIT-3 (Wang et al., 2022) further enhances the MoE part for multiple tasks jointly. BLIP-2 (Li et al., 2023) adopts fixed pre-trained modality models for efficient training with Q-former.

BEiT (Bao et al., 2021) and MAE (He et al., 2022) were proposed in computer vision based on an encoder-decoder framework and the masked image modeling task. These two works brought a new perspective on vision pre-training. However, it still needs further explorations to conduct V&L pre-training using the autoencoder framework. In this paper, we propose LightVLP, a new framework for V&L pre-training, improving both performance and training efficiency at the same time.

3. Our Approach

In this section, we present the proposed LightVLP model. As shown in Fig.2, LightVLP is composed of two symmetrical autoencoders, one for image processing and the other for text representations. The encoders of LightVLP take the masked image patches and text tokens as input and conduct uni-modal representations as well as cross-modal interactions. In the decoder, we reconstruct the masked image pixels and text tokens. LightVLP comprises five major components: masking and removing in the input layer, contrastive learning in the encoders, gated interaction mechanism in the encoders, image and text reconstructions in the decoders as well as several other cross-modal training tasks.

3.1. Input Masking and Removing

The input is an image $I_x \in \mathbb{R}^{H\times W^0\times C}$ and the corresponding text $W = \{w_1, w_2, ..., w_N\}$, where the image resolution is (H, W^0) and C is the number of channels. We follow MAE (He et al., 2022) to divide the input image into several patches of resolution (P, P). The image patches are $X = \{x_1, x_2, ..., x_M\}$, where $x_i \in \mathbb{R}^{P^2\times C}$. Then we randomly mask and remove some patches with an image masking rate p_{xmsk}. We take the unmasked patches as the image input of LightVLP, i.e., $\tilde{X} = \{x_1, x_2, ..., x_{\tilde{M}}\}$, where $\tilde{M} = M \times (1 - p_{xmsk})$. As for the text, we first segment the input text using BERT (Devlin et al., 2018) tokenizer. Then, we randomly mask some tokens with a masking rate, i.e., p_{wmsk}, and feed the remaining ones to LightVLP, i.e., $\tilde{W} = \{w_1, w_2, ..., w_{\tilde{N}}\}$, where

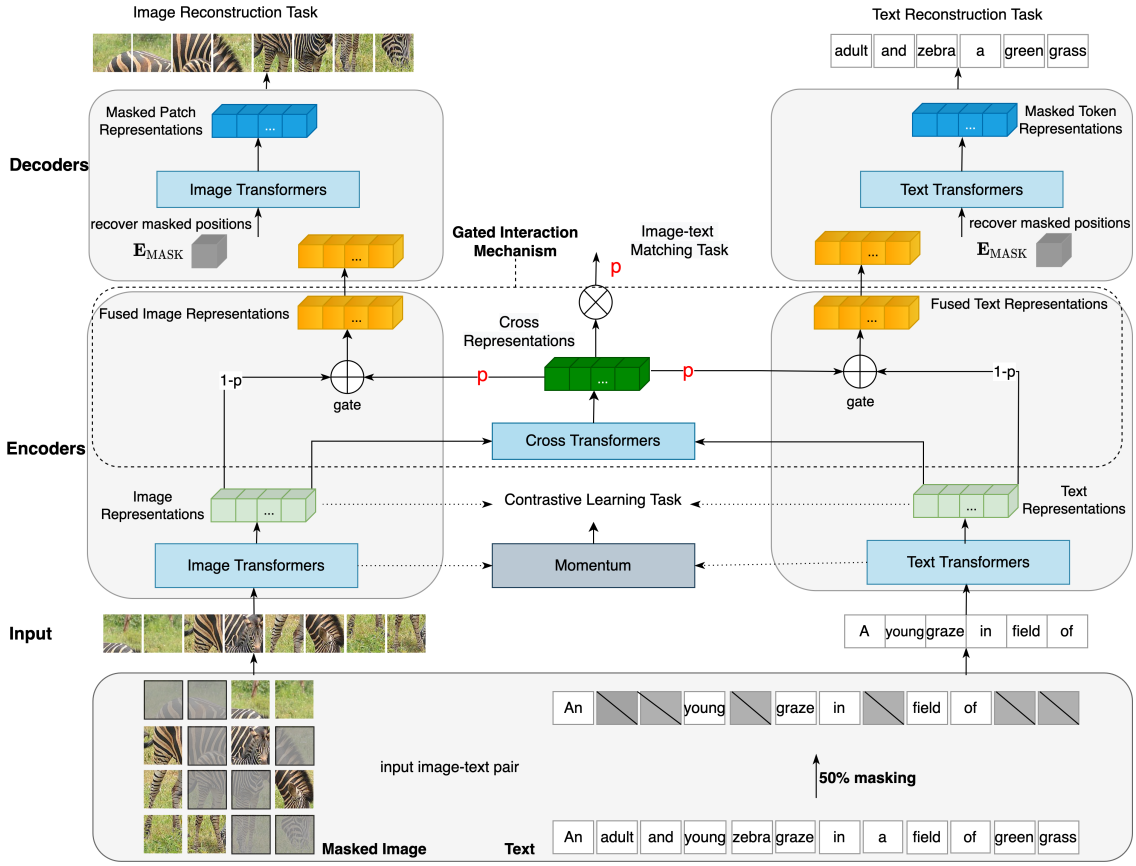

Figure 2: The architecture of LightVLP, which comprises two symmetrical and interactive autoencoders, one for image and the other for text. In this case, 50% image and text tokens are masked and removed in the input. Besides, the two encoders exchange information via contrastive learning and cross Transformers. In the gated interaction mechanism, the gates choose between single-modal and cross-modal representations to reduce noise from unmatched image-text pairs. The decoders aim to reconstruct the masked tokens. $\mathbf{E}_{[\text{MASK}]}$ indicates the [MASK] embedding.

$\tilde{N} = N \times (1 - p_{wmsk})$. Randomly masking and removing a large portion of input tokens has the following two advantages: On the one hand, the input sequences of Transformers are shortened, which can reduce the computational cost significantly. On the other hand, the tasks become more challenging for grasping better understanding abilities, which has been verified in uni-modal pre-training, for example, Wettig et al. (2022) and He et al. (2022).

3.2. Lightweight Encoders

In Fig.2, the two encoders are of the same structure, each involving the uni-modal encoding, contrastive learning and gated interaction mechanism.

Uni-modal Encoding. In uni-modal encoding, we apply two separate uni-modal Transformers of l_e layers and hidden size H_e to get the deep representations for the image and text, respectively,

$$\begin{aligned}\mathbf{H}^{l+1} = f_{\text{LN}}\Big(f_{\text{LN}}(\mathbf{H}^l + f^l_{\text{SA}}(\mathbf{H}^l)) + \\ f^l_{\text{FF}}(f_{\text{LN}}(\mathbf{H}^l + f^l_{\text{SA}}(\mathbf{H}^l)))\Big),\end{aligned} \tag{1}$$

where f_{LN} indicates the layer normalization operation. f_{FF} is the feed forward layer involving two fully-connected sub-layers. f_{SA} is the self-attention layer which uses multi-head attention to grasp token-level relationships. After the uni-modal encoding, we can get uni-modal representations for the remaining image patches and text tokens,

$$\begin{aligned}\mathbf{H}_x{}^{l_e} &= \mathbf{H}_{x0}{}^{l_e}, \mathbf{H}_{x1}{}^{l_e}, ..., \mathbf{H}_{x\tilde{M}}{}^{l_e}; \\ \mathbf{H}_w{}^{l_e} &= \mathbf{H}_{w0}{}^{l_e}, \mathbf{H}_{w1}{}^{l_e}, ..., \mathbf{H}_{w\tilde{N}}{}^{l_e},\end{aligned} \tag{2}$$

where l_e indicates the last hidden layer. $\mathbf{H}_x{}^{l_e}$ and $\mathbf{H}_w{}^{l_e}$ are the representations for the input image and text, respectively. $\mathbf{H}_{*0}$ are hidden states corresponding to the [CLS] positions.

Contrastive Learning. We use the uni-modal representations to conduct contrastive learning. In LightVLP, contrastive learning serves as an efficient way to learn shallow cross-modal interactions, by involving large amounts of negative samples for computation.

Specifically, we apply linear transformations, $f_x(\cdot)$ and $f_w(\cdot)$ to reduce the dimension of the

previous uni-modal representations to 256-d embeddings. Then we normalize the representations for dot similarity calculation as $\mathbf{I} = f_{\text{NORM}}(f_x(\mathbf{H}_{x_0}^{l_e})), \mathbf{T} = f_{\text{NORM}}(f_w(\mathbf{H}_{w0}{}^{l_e}))$. We follow the momentum method by MoCo (He et al., 2020). In order to introduce more negative samples, we introduce two queues to keep the previous M_c image and text representations from the momentum uni-modal encoders respectively. For the current input image and text, we compute their similarity with text and image representations in the corresponding queues by dot production and then normalize the similarity scores as follows,

$$\begin{aligned} p_k^{i2t}(\mathbf{I}) &= \frac{\exp(\mathbf{I} \cdot \mathbf{T}_k{}'/\kappa)}{\sum_{j=1}^{M_c} \exp(\mathbf{I} \cdot \mathbf{T}_j{}'/\kappa)}; \\ p_k^{t2i}(\mathbf{T}) &= \frac{\exp(\mathbf{T} \cdot \mathbf{I}_k{}'/\kappa)}{\sum_{j=1}^{M_c} \exp(\mathbf{T} \cdot \mathbf{I}_j{}'/\kappa)}, \end{aligned} \tag{3}$$

where κ indicates the temperature scalar. $\mathbf{I}'_*$ and $\mathbf{T}'_*$ indicate the image and text representations from the queues.

Gated Interaction Mechanism (GIM).

Existing hybrid models usually fuse image-text representations using cross attention mechanism. Then, they directly take the fused representations for masked language modeling or masked image modeling tasks. However, most image-text pairs are web noisy data, which means there are unrelated image-text pairs in the corpora. It is obvious that fused representations of unrelated image-text pairs are worse compared to uni-modal representations for reconstruction tasks. To this end, we propose the gated interaction mechanism to tackle noise in image-text pairs. As for a matched image-text pair, the model prefers cross-modal representations as the output of the encoders, which will be passed to decoders for reconstruction tasks. By contrast, for an unmatched image-text pair, the encoder is more likely to output uni-modal representations.

The gated interaction mechanism is depicted in Fig.2. First, in order to fuse information from the cross-modal input, we apply l_c cross Transformers to grasp cross-modal interactions. The input is the uni-modal representations. We obtain the cross representations as the following equations.

$$\begin{aligned} \mathbf{S}_x^{l+1} = f_{\text{LN}}\Big(f_{\text{LN}}\big(\mathbf{S}_x^l + f_{\text{CROSS}}^l(\mathbf{S}_x^l|\mathbf{S}_w^l)\big) + \\ f_{\text{FF}}^l\big(f_{\text{LN}}(\mathbf{S}_x^l + f_{\text{CROSS}}^l(\mathbf{S}_x^l|\mathbf{S}_w^l))\big)\Big), \\ f_{\text{CROSS}}(\mathbf{S}_x^l|\mathbf{S}_w^l) = \text{softmax}(\mathbf{Q}_x \cdot \mathbf{K}_w/\sqrt{d_k})\mathbf{V}_w, \end{aligned} \tag{4}$$

where f_{LN} is layer normalization. The feed-forward sub-layer f_{FF} involves two dense connected sub-layers. f_{CROSS} is the cross-modal attention mechanism. $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value calculated by linear mapping layers. d_k is the hidden dimension. We ignore the equation of multiple heads for simplicity.

Next, we calculate the image-text matching probability by applying a linear network f_{GATE} and sigmoid function $\sigma(\cdot)$ as,

$$p = \sigma\big(f_{\text{GATE}}(\mathbf{S}_{x_0}^{l_c})\big), \tag{5}$$

where p can be regarded as a soft gate to choose between cross-modal representations and uni-modal representations. A higher p indicates that the image and text match well, and thus the cross-modal representations should be more considered in the following tasks. We use the gate to merge uni-modal and cross-modal representations,

$$\mathbf{O} = (1-p) \cdot \mathbf{H}^{l_e} + p \cdot \mathbf{S}^{l_c}, \tag{6}$$

where $\mathbf{O}$ is the output of encoders which serves as part of input for the decoders. If the image-text inputs are matched, the cross representations are more likely passed to decoders for reconstruction. While the image-text pair is unmatched, most uni-modal representations are passed to decoders.

3.3. Lightweight Decoders

The decoders are l_d successive layers of Transformers with hidden H_d. Compared to encoders, the decoders are lightweight for the purpose of efficiency. Specifically, their input is the output of encoders $\mathbf{O}$ after recovering the masked positions with [MASK] embedding $\mathbf{E}_{\text{[MASK]}}$. We adopt Transformers to interact with the masked positions and remaining tokens. After that, we make predictions for the masked positions. The decoders are designed to resolve two similar tasks for two modalities: the masked image reconstruction task and the masked token reconstruction task. As for the masked image reconstruction task, we aim to reconstruct the pixels of the original image. As for the masked token reconstruction task, we reconstruct texts by predicting token IDs.

3.4. Training Objectives

We design four tasks to pre-train LightVLP, including Masked Image Reconstruction (MIR), Masked Token Reconstruction (MTR), Image Text Contrastive (ITC) learning and Image Text Matching (ITM). The final loss is the sum of the four task losses as $\mathcal{L} = \mathcal{L}_{\text{MIR}} + \mathcal{L}_{\text{MTR}} + \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}}$ (we empirically set the loss weights equally).

Masked Image Reconstruction. MIR is designed to reconstruct the masked patches by predicting the pixels. The output of the image decoder involves the predicted pixels for each patch. We apply mean square error as the loss function to

measure the distance of the reconstructed pixels and the ground-truth values,

$$\mathcal{L}_{\text{MIR}}(\Theta) = \frac{1}{|M_x|} \sum_{m \in M_x} f_{\text{MSE}}(\mathbf{x}_m, \hat{\mathbf{x}_m}), \quad (7)$$

where M_x is the set of masked image tokens. f_{MSE} is the function to compute the mean square error. $\hat{\mathbf{x}_m}$ indicates the predicted pixels while $\mathbf{x}_m$ represents the ground-truth values.

Masked Token Reconstruction. MTR is similar to the masked language modeling task in BERT. The key difference lies in two perspectives. First, the masked tokens are removed for the input and the decoder reserves the masked positions for predictions. Second, the masking rate is larger compared to BERT to make the task tougher as well as reduce the computational complexity. The loss function is computed as the following equation,

$$\mathcal{L}_{\text{MTR}}(\Theta) = \frac{1}{|M_w|} \sum_{m \in M_w} f_{\text{CE}}(\mathbf{y}_m, \hat{\mathbf{y}}_m), \quad (8)$$

where M_w is the set of masked tokens. $\hat{\mathbf{y}_m}$ indicates the predictions while $\mathbf{y}_m$ represents the one-hot ground-truth. f_{CE} represents cross entropy.

Image Text Contrastive Learning. We learn V&L representations through V&L contrastive learning. Specifically, the model is designed to differentiate the positive image-text samples from the negative ones. Suppose that $\mathbf{y}^{i2t}(\mathbf{I})$ and $\mathbf{y}^{t2i}(\mathbf{T})$ are the ground-truth one-hot labels indicating the image-text relationships. The image-text contrastive loss can be calculated by the cross entropy function f_{CE} using the predictions ($\mathbf{p}^{i2t}(\mathbf{I})$, $\mathbf{p}^{t2i}(\mathbf{T})$) and the ground-truth in the way as,

$$\begin{aligned}\mathcal{L}_{\text{ITC}}(\Theta) = &\frac{1}{2} f_{\text{CE}}\big(\mathbf{y}^{i2t}(\mathbf{I}), \mathbf{p}^{i2t}(\mathbf{I})\big) + \\ &\frac{1}{2} f_{\text{CE}}\big(\mathbf{y}^{t2i}(\mathbf{T}), \mathbf{p}^{t2i}(\mathbf{T})\big).\end{aligned} \quad (9)$$

Image Text Matching. ITM is to predict whether the image-text pairs are matched or unmatched. Compared to ITC, ITM is performed after the gated interaction mechanism. We directly take in the image-text matching probability from Eq.(5) and compute the loss function as,

$$\mathcal{L}_{\text{ITM}} = f_{\text{CE}}\big(\mathbf{y}_{itm}, [1-p,\ p]\big), \quad (10)$$

where $\mathbf{y}_{itm}$ is a 2-dimensional vector indicating the ground-truth labels.

In ITM, we also introduce hard negatives by in-batch hard negative sampling. Specifically, we first calculate the in-batch normalized matching probabilities using Eq.(3). Next, we take these matching probabilities to sample more challenging image-text negatives. In implementation, we sample two hard negative pairs for each input image-text pair, of which loss can be computed by Eq.(10).

4. Experiments

4.1. Experimental Settings

Pre-training Corpus. We followed ALBEF (Li et al., 2021) and collected the publicly available pre-training datasets from the Internet, including SBU (Ordonez et al., 2011), MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017) and Conceptual Captions (Sharma et al., 2018). The final training dataset contains about 4 million images and 5 million image-text pairs. Involving more training data could induce more significant improvements, which is verified by Li et al. (2022) and Li et al. (2021). For fairness, we only perform model comparisons under similar amounts of training corpora (about 4 million images).

Parameter Settings. We pre-trained two models, LightVLP$_{base}$ and LightVLP$_{large}$ model. LightVLP$_{large}$ has 805M trainable parameters but its FLOPs is similar to ALBEF (see Table 1). LightVLP$_{base}$ has similar parameters as ALBEF, i.e., 433M (LightVLP$_{base}$) and 420M (ALBEF). As for LightVLP$_{large}$, the hidden layer and dimension of uni-modal Transformers were 12 and 1024, while the hidden layer and dimension of decoders were set to 4 and 512, respectively. Training batch size was set to 448 under the 50% image and 50% text masking setting. For LightVLP$_{base}$ model, we set the hidden layer and dimension of uni-modal Transformers to 12 and 768, while setting the values of decoders to 4 and 256, respectively. The cross Transformers had 4 layers, of which dimensions were set the same as the uni-modal Transformers. The training batch size was set to 512. As for contrastive learning, queue size M_c was set to 655,36 and κ was set to 0.05. We took images of 256×256 as input during pre-training and kept this setting in the fine-tuning. We set patch resolution to 16×16. Before masking, the maximum length of the original patch was 256 and the max text tokens were 32. If masking rates were set to 50% and 50%, the input sequence length of encoders is 144 in total $(256 \times 50\% + 32 \times 50\%)$. We used 8 NVIDIA A100 to train each model. We applied Adafactor optimizer with a learning rate of 1e-4 and weight decay of 0.02.

4.2. Benchmarks and Main Results

We pre-train LightVLP under 50% and 50% masking setting and evaluate it on six benchmarks. In this section, we describe the results one by one.

Visual Retrieval Task. We mainly evaluate the Visual Retrieval task on COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015), which are two commonly used cross-modal retrieval datasets.

The model performance is summarized in Table 1. From Table 1, we have the following ob-

Models	GFLOPs	MS COCO						Flickr30K					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER(Chen et al., 2020)	-	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA(Gan et al., 2020)	-	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR(Li et al., 2020c)	-	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO(Li et al., 2020b)	-	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
ALBEF(Li et al., 2021)	135.9	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	**82.8**	**96.7**	98.4
LightVLP$_{base}$	**75.6**	73.5	91.7	96.1	56.5	81.4	88.8	94.1	99.5	**99.8**	81.1	96.1	97.8
LightVLP$_{large}$	124.4	**75.3**	**93.2**	**96.6**	**57.6**	**82.2**	**89.4**	**94.4**	**99.6**	99.7	82.5	96.5	**98.6**

Table 1: Results on Image Retrieval and Text Retrieval tasks of MS COCO and Flickr30K datasets. We report the R@1, R@5 and R@10 results. TR/IR represent the text retrieval/image retrieval tasks.

servations. (1) Compared to state-of-the-art ALBEF, LightVLP$_{base}$ model behaves competitively by saving about 44% computation evaluated by FLOPs. (2) LightVLP$_{large}$ performs much better than the LightVLP$_{base}$ model. It achieves an improvement of 1.1% by R Mean (averaged value of R@1, R@5 and R@10) on COCO. (3) On COCO, LightVLP$_{large}$ is 1.1% more accurate than the previous ALBEF and is much better than other baselines. In particular, it improves R@1 of text retrieval by 2.2%. (4) As for the Flickr30K dataset, the improvements are not so obvious. That is because the results of these tasks are near the upper bound of models, with R@10 valued near 100%.

In the following part, we use LightVLP$_{large}$ to conduct further comparisons and analyses.

Visual Grounding Task. The task aims to compute a region from the input image corresponding to the input text. One commonly used dataset is RefCOCO+ (Yu et al., 2016). We test LightVLP on RefCOCO+ and tabularize the result in Table 2. Our model shows an improvement of 1.16% on the TestA set and an improvement of 1.08% on TestB.

Models	Val	TestA	TestB
ARN(Liu et al., 2019a)	32.78	34.35	32.13
CCL(Zhang et al., 2020)	34.29	36.91	33.56
ALBEF(Li et al., 2021)	58.46	65.89	46.25
LightVLP	**58.60**	**67.05**	**47.33**

Table 2: Results on the RefCOCO+ dataset.

Visual Reasoning Task. Given a piece of text and a pair of images, the model should predict whether the text describes the two images or not. We employ NLVR^2 (Suhr et al., 2018) to evaluate our model on this task and list the results in Table 3. The improvement is 1.0% on the test set, which indicates our model could solve more challenging tasks such as the Visual Reasoning task.

Visual Entailment Task. The task is to reason the relationship between an image and a piece of text, i.e., entailment, neutral and contradiction. We use SNLI-VE (Xie et al., 2019) to test the ability of our model on the Visual Entailment task. As Table 3 shows, LightVLP performs over 3.1% better than OSCAR and UNITER models while achieving an improvement of 1.1% compared with ALBEF.

Models	NLVR^2		SNLI-VE	
	dev	test	dev	test
VisualBERT(Li et al., 2019)	67.40	67.00	-	-
LXMERT(Tan and Bansal, 2019)	74.90	74.50	-	-
12-in-1(Lu et al., 2020)	-	78.87	-	76.95
UNITER(Chen et al., 2020)	77.18	77.85	78.59	78.28
VL-BART(Cho et al., 2021)	-	73.60	-	-
ViLT(Kim et al., 2021)	75.24	76.21	-	-
OSCAR(Li et al., 2020c)	78.07	78.36	-	-
VILLA(Gan et al., 2020)	78.39	79.30	79.47	79.03
ALBEF(Li et al., 2021)	80.24	80.50	80.14	80.30
LightVLP	**81.27**	**81.53**	**81.66**	**81.40**

Table 3: Results on NLVR^2 and SNLI-VE datasets.

Visual Question Answering (VQA). VQA (Goyal et al., 2017) takes an image and a question as inputs and outputs the answer for the question from limited candidates. We treat this task as a classification task instead of a generation task. From Table 4, our model is nearly 1.6% better than strong baselines, showing the superiority of our framework and proposed mechanisms.

Models	test-dev	test-std
VisualBERT (Li et al., 2019)	70.80	71.00
LXMERT (Tan and Bansal, 2019)	72.42	72.54
12-in-1 (Lu et al., 2020)	73.15	-
UNITER (Chen et al., 2020)	72.70	72.91
VL-BART (Cho et al., 2021)	-	71.30
ViLT (Kim et al., 2021)	70.94	-
OSCAR (Li et al., 2020c)	73.16	73.44
VILLA (Gan et al., 2020)	73.59	73.67
ALBEF (Li et al., 2021)	74.54	74.70
LightVLP	**76.19**	**76.30**

Table 4: Results on the VQA dataset.

4.3. Masking Rate Analysis

We further train the model under different image and text token masking rates to compare the corresponding model performance and computational cost in terms of FLOPs. The experiments are conducted on Text Retrieval and Image Retrieval tasks using the MS COCO dataset. From Table 5, we have the following observations: (1) When we set

Masking ratio image	Masking ratio text	GFLOPs	TR R@1	TR R@5	TR R@10	TR R Mean	IR R@1	IR R@5	IR R@10	IR R Mean	Overall R Mean
25%	25%	168.0	74.90	92.54	96.52	87.99	57.49	**82.49**	**89.38**	**76.45**	82.22
25%	50%	165.6	74.80	92.74	96.48	88.01	57.02	82.11	**89.38**	76.17	82.09
25%	75%	163.2	74.10	92.78	96.24	87.71	55.95	81.30	88.66	75.30	81.51
50%	25%	126.9	74.66	92.32	96.30	87.76	57.25	81.92	89.06	76.07	81.92
50%	50%	124.4	**75.30**	**93.18**	**96.64**	**88.37**	**57.63**	82.24	**89.38**	76.42	**82.39**
50%	75%	122.0	74.46	92.02	96.14	87.54	56.12	81.31	88.97	75.47	81.50
75%	25%	85.7	74.20	92.72	96.30	87.74	56.63	81.61	89.13	75.79	81.77
75%	50%	83.3	73.98	92.16	95.84	87.32	56.04	81.32	88.53	75.29	81.31
75%	75%	**80.9**	72.74	92.10	95.92	86.92	54.64	80.50	88.21	74.45	80.69

Table 5: Model performance and computational cost under different masking ratios. R Mean indicates the averaged value of R@1, R@5, R@10. Masking ratio (50%, 50%) achieves the overall best performance.

masking rates of image patch and text token to 25% and 25% respectively, the model behaves well with an overall R Mean 82.22%. The corresponding GFOLPs becomes as large as 168.0; (2) The 50% and 50% setting achieves the best result with an overall R Mean of 82.39%, and similar conclusions are also found in other tasks. Therefore, considering the balance of computational cost and model performance, it is a mostly recommended setting; (3) When the masking rates are set to 75% and 75%, GFLOPs reaches the lowest. But it tends to cause loss of effect; (4) Masking more images reduces the GFLOPs more significantly. That is because the maximum length of the original patch is 256 while the max text token length is 32.

From our perspective, increasing the masking rates will have the following influences. On the one hand, it increases the difficulty of reconstruction tasks. Making the tasks more challenging can benefit the model. On the other hand, it will reduce the input tokens by removing the masked ones, which harms the model performance under the same amount of training data. These two factors may cause performance fluctuations under different masking rates.

4.4. Ablation Study

Gated interaction mechanism. To evaluate the relative improvement of the proposed gated interaction mechanism. We make an ablation study by disabling the gate in this mechanism of LightVLP. In other words, we directly take the cross representations as the output of the encoder by forcing $p = 1$ as shown in Fig. 2. We use the same model settings to pre-train the model and compare its performance with LightVLP on Image Retrieval and Text Retrieval tasks using the MSCOCO dataset.

From Table 6, the gate of our gated interaction mechanism makes an improvement of 0.57% by R Mean compared to LightVLP w/o gate. Especially, it improves 1.46% by R@1 on text retrieval. Since the training datasets of LightVLP are mostly collected from the web, there exists much noise. By introducing the gated interaction mechanism, the gate enables the model to selectively union cross-modal and single-modal representations according to the image-text matching probability. As for a matched image-text pair, the model tends to apply cross-modal representations for reconstructions. By contrast, for an unmatched pair, the model conducts reconstructions mainly using uni-modal representations. GIM significantly reduces the effect of unmatched image-text pairs in training, which is the key reason for the improvement.

Models	TR R@1	TR R@5	TR R@10	IR R@1	IR R@5	IR R@10
LightVLP	**75.30**	**93.18**	**96.64**	**57.63**	**82.24**	**89.38**
w/o gate	73.84	92.94	96.50	56.85	81.83	89.04

Table 6: Ablation study on MSCOCO (TR/IR).

Training objectives. We further conduct several ablation versions to prove the effectiveness of different pre-training tasks. Experimental results are recorded in Table 7. We find that all pre-training tasks contribute to the performance on SNLI-V, NLVR^2, and VQA, among which ITC and ITM show more gains than MIR.

Models	SNLI-VE	NLVR^2	VQA
LightVLP	**81.40**	**81.53**	**76.19**
LightVLP w/o gate	80.27	80.87	73.34
LightVLP w/o MIR	81.05	80.60	74.91
LightVLP w/o ITC	80.33	79.51	74.17
LightVLP w/o ITM	79.77	80.40	74.04

Table 7: Ablation study on SNLI-VE, NLVR^2, VQA.

4.5. In-depth Analysis on GIM

To verify that GIM works as we designed, we study the distribution of p in Eq. (5) in the pre-training dataset. From Fig. 4 we discover that: over 20% image-text pairs are considered unrelated ($p < 0.2$) and more than 50% image-text pairs are partially matched ($p < 0.8$), which indicates there exists much noise in real-world pre-training corpora.

Further, we conduct a case study by sampling some image-text pairs with high/low p scores listed

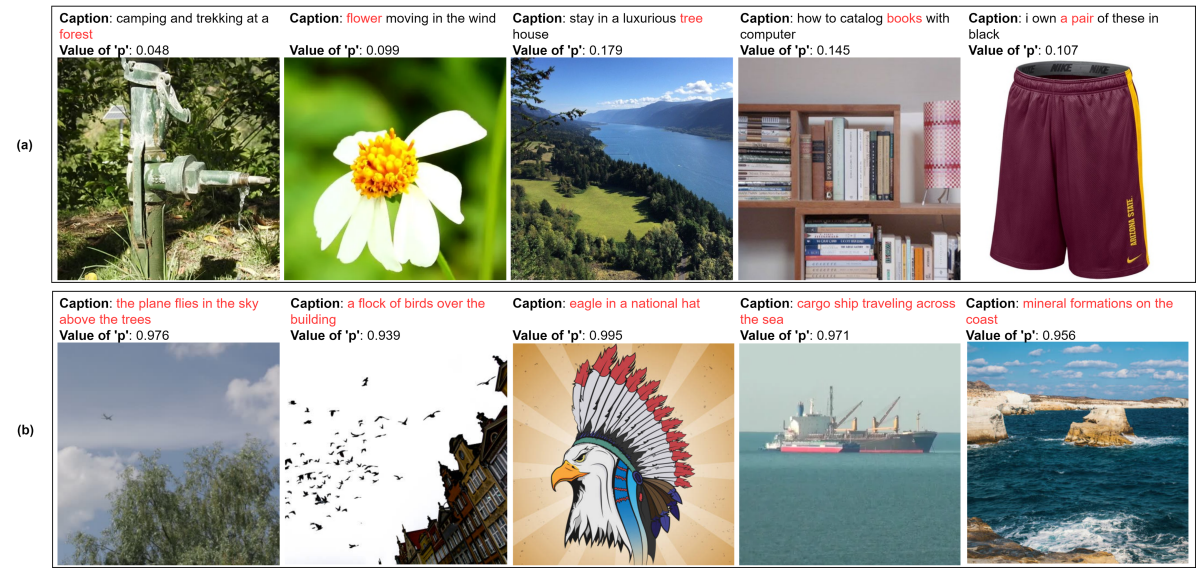

Figure 3: Matched (b) and unmatched (a) cases with different p in GIM. GIM functions as expected.

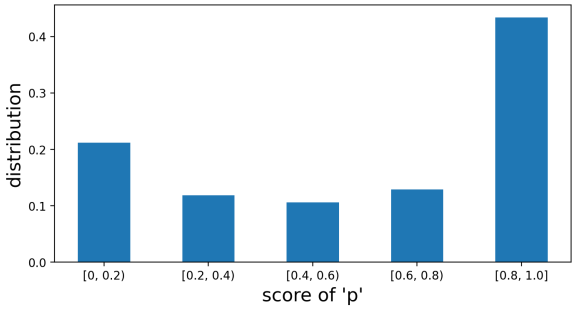

Figure 4: Distribution of matching probability p.

in Fig. 3, including 5 matched (b) and 5 unmatched (a) cases. As for matched image-text pairs (b), LightVLP tends to give higher p to apply cross-modal representations for reconstructions. By contrast, for unmatched pairs (a), LightVLP conducts reconstructions mainly using uni-modal representations instructed by lower p. In this way, it can reduce the side effects of noisy data in pre-training.

4.6. Discussions and Limitations

In this section, we analyze the limitations of our LightVLP considering the main challenges in V&L pre-training as described in the Introduction: First, the pre-training complexity is still high. Although reducing the input length by a large margin in the encoders can improve the training efficiency significantly (44% by FLOPs), it is still training expensive. That is because the fundamental cause of high computational costs lies in the Transformer itself. Hence, improving the basic structures of Transformers can benefit more. Second, the self-supervised tasks need to be improved. In LightVLP, we introduce four tasks to improve the V&L pre-training. Two main tasks are masked image modeling and masked token modeling. We increase the masking rates to make these tasks tougher for better V&L understanding ability. It has two disadvantages: (1) Increasing the masking ratios will decrease the input tokens of the encoders, which harms the model performance. (2) Keeping the masking ratios constant means keeping the task difficulty unchanged, which is against the order of human learning, i.e., from easy to hard. We will further improve the model via curriculum learning. Third, as for the noisy data, GIM is effective in dealing with noise from the modeling aspect. However, the noisy image-text pairs still exist in the training data and harm the model performance. More purified image-text pre-training corpora are needed.

5. Conclusions and Future Work

In this paper, we propose LightVLP, a new framework for V&L pre-training, in which we mask and remove large portions of input tokens as the encoders' input, reducing computational complexity significantly. We explore different settings of masking ratios and propose a gated interaction mechanism to automatically choose between cross-modal and uni-modal representations for reconstruction, which helps reduce noise from unmatched image-text pairs. Our models show competitive results on six commonly used datasets.

In the future, we will evaluate more lightweight methods to make full use of the powerful large pre-trained models without much cost. We will also explore the practical usage of lightweight VLP models in the era of LLMs considering both benefits and costs, efficiently and economically fine-tuning VLP models for different downstream tasks.

Acknowledgements

This paper is supported by the Science and Technology Development Fund of Macau SAR (File no. 0081/2022/A2, 0123/2022/AFJ, and 0015/2019/AKP), GuangDong Basic and Applied Basic Research Foundation (No. 2020B1515130004), and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Proceedings of NeurIPS*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. 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, pages 11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020b. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. 2019a. Adaptive reconstruction network for weakly supervised referring expression grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2611–2620.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers; distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134.