VLMO: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a unified **V**ision-**L**anguage pretrained **Mo**del (**VLMO**) that jointly learns a dual encoder and a fusion encoder with a modular Transformer network. Specifically, we introduce **M**ixture-**o**f-**M**odality-**E**xperts (**MOME**) Transformer, where each block contains a pool of modality-specific experts and a shared self-attention layer. Because of the modeling flexibility of MOME, pretrained VLMO can be fine-tuned as a fusion encoder for vision-language classification tasks, or used as a dual encoder for efficient image-text retrieval. Moreover, we propose a stagewise pre-training strategy, which effectively leverages large-scale image-only and text-only data besides image-text pairs. Experimental results show that VLMO achieves state-of-the-art results on various vision-language tasks, including VQA, NLVR2 and image-text retrieval. The code and pretrained models are available at https://github.com/xxx/yyy.

1 Introduction

Vision-Language (VL) pre-training [30, 40, 35, 26, 20, 23] learns generic cross-modal representations from large-scale image-text pairs. Previous models usually employ image-text matching, image-text contrastive learning, masked region classification/feature regression, word-region/patch alignment and masked language modeling to aggregate and align visual and linguistic information. Then the pretrained models can be directly fine-tuned on downstream vision-language tasks, such as VL retrieval and classification (visual question answering, visual reasoning, etc.).

Two mainstream architectures are widely used in previous work. CLIP [35] and ALIGN [18] adopt a *dual-encoder* architecture to encode images and text separately. Modality interaction is handled by the cosine similarity of the image and text feature vectors. The dual-encoder architecture is effective for retrieval tasks, especially for masses of images and text. Feature vectors of images and text can be pre-computed and stored. However, the shallow interaction between images and text is not enough to handle complex VL classification tasks. ViLT [20] finds that CLIP gives a relatively low accuracy on visual reasoning task. Another line of work [30, 40, 42, 3, 20, 23] relies on a fusion encoder with cross-modal attention to model image-text pairs. Multi-layer Transformer [44] networks are usually employed to fuse image and text representations. The *fusion-encoder* architecture achieves superior performance on VL classification tasks. But it requires to jointly encode all possible image-text pairs to compute similarity scores for retrieval tasks. The quadratic time complexity leads to a much slower inference speed than the dual-encoder models whose time complexity is linear.

In order to take advantage of the two types of architectures, we propose a unified **V**ision-**L**anguage pretrained **Mo**del (**VLMO**) that can be used as either a dual encoder to separately encode images and text for retrieval tasks, or used as a fusion encoder to model the deep interaction of image-text pairs for classification tasks. This is achieved by introducing **M**ixture-**o**f-**M**odality-**E**xperts (**MOME**) Transformer that can encode various modalities (images, text, and image-text pairs) within

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

a Transformer block. MoME employs a pool of modality experts to replace the feed-forward network in standard Transformer. It captures modality-specific information by switching to different modality experts, and uses the shared self-attention across modalities to align visual and linguistic information. Specifically, MoME Transformer consists of three modality experts, namely vision expert for image encoding, language expert for text encoding, and vision-language expert for image-text fusion. Thanks to the modeling flexibility, we can reuse MoME Transformer with the shared parameters for different purposes, i.e., text-only encoder, image-only encoder, and image-text fusion encoder.

VLMo is jointly learned with three pre-training tasks, namely image-text contrastive learning, image-text matching, and masked language modeling. In addition, we propose a stagewise pre-training strategy to effectively leverage large-scale image-only and text-only corpus besides image-text pairs in VLMo pre-training. We first pretrain vision experts and self-attention modules of MoME Transformer on image-only data using masked image modeling proposed in BEiT [2]. We then pretrain language experts on text-only data using masked language modeling [10]. Finally, the model is used to initialize vision-language pre-training. By getting rid of the limited size of image-text pairs and their simple and short captions, stagewise pre-training on large amounts of image-only and text-only data helps VLMo to learn more generalizable representations.

Experimental results demonstrate that VLMo achieves state-of-the-art results on vision-language retrieval and classification tasks. Our model, used as a dual encoder, outperforms fusion-encoder-based models [3, 14, 20, 23] while enjoying a much faster inference speed on retrieval tasks. Moreover, our model also achieves state-of-the-art results on visual question answering (VQA) and natural language for visual reasoning (NLVR2), where VLMo is used as a fusion encoder.

Our main contributions are summarized as follows:

- We propose a unified vision-language pretrained model VLMo that can be used as a fusion encoder for classification tasks, or fine-tuned as a dual encoder for retrieval tasks.
- We introduce a general-purpose multimodal Transformer for vision-language tasks, namely MoME Transformer, to encode different modalities. It captures modality-specific information by modality experts, and aligns contents of different modalities by the self-attention module shared across modalities.
- We show that stagewise pre-training using large amounts of image-only and text-only data greatly improves our vision-language pretrained model.

2 Related Work

Pre-training with Transformer [44] backbone networks has substantially advanced the state of the art across natural language processing [34, 10, 28, 22, 11, 36, 1, 7, 8, 4–6, 31], computer vision [12, 43, 2] and vision-language [42, 40, 3, 48, 35, 18, 20, 23] tasks.

The approaches of vision-language pre-training can be divided into two categories. The first category utilizes a dual encoder to encode images and text separately, and uses cosine similarity or a linear projection layer to model the interaction between images and text [35, 18]. Image-text contrastive learning is usually employed to optimize the model. Dual-encoder models are effective for vision-language retrieval tasks. However, the simple interaction is not enough to handle tasks that require complex reasoning, such as visual reasoning and visual question answering (VL classification tasks). The second category models the interaction of images and text using a deep fusion encoder with cross-modal attention [42, 30, 40, 24, 49, 3, 26, 25, 14, 48, 16, 17, 20, 23, 45]. Image-text matching, masked language modeling, word-region/patch alignment, masked region classification and feature regression are widely used to train fusion-encoder-based models. These models achieve better performance for vision-language classification tasks, while the joint encoding of all image-text pairs leads to a slow inference speed for retrieval tasks. A large portion of fusion-encoder-based models rely on an off-the-shelf object detector like Faster R-CNN [37] to obtain image region features. Generating region features slows down the inference speed and renders the approach less scalable. Recently, Pixel-BERT [16] removes object detector and encodes images into grid features by convolutional neural networks. ALBEF [23] employs image Transformer [12, 43] to obtain the representations of images, and uses text Transformer [10] to learn the contextualized representations of text. These representations are then fused by cross-modal attention. ViLT [20] encodes images into patch embeddings, and then feed the concatenation of image patch embeddings and word embeddings

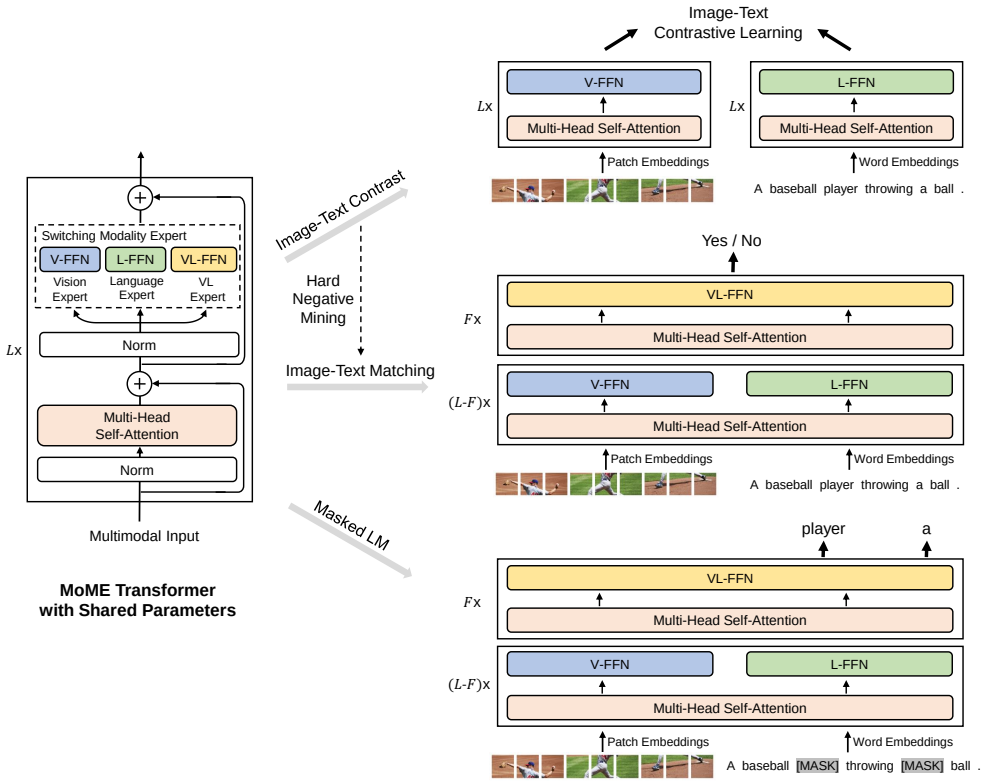


Figure 1: Overview of VLMO pre-training. We introduce mixture-of-modality-experts (MOME) Transformer to encode different modality input by modality-specific experts. The model parameters are shared across image-text contrastive learning, masked language modeling, and image-text matching pre-training tasks. During fine-tuning, the flexible modeling enables us to use VLMO as either a dual encoder (i.e., separately encode images and text for retrieval tasks) or a fusion encoder (i.e., jointly encode image-text pairs for better interaction across modalities).

into a Transformer network to learn contextualized representations and model the interaction of images and text.

Different from previous work, our unified pre-training using shared MOME Transformer enables the model perform separate encoding for retrieval tasks, and jointly encode image-text pairs to capture deeper interaction for classification tasks. Our model achieves competitive performance, while enjoying a faster inference speed for both retrieval and classification tasks.

3 Methods

Given image-text pairs, VLMO obtains image-only, text-only and image-text pair representations by the MOME Transformer network. As shown in Figure 1, the unified pre-training optimizes shared MOME Transformer with image-text contrastive learning on image-only and text-only representations, image-text matching and masked language modeling on image-text pair representations. Thanks to the modeling flexibility, the model can be used as a dual encoder for retrieval tasks to encode images and text separately during fine-tuning. It can also be fine-tuned as a fusion encoder to model deeper modality interaction of images and text for classification tasks.

3.1 Input Representations

Given an image-text pair, we encode the pair into image, text and image-text vector representations. These representations are then fed into the MOME Transformer to learn contextualized representations and align image and text feature vectors.

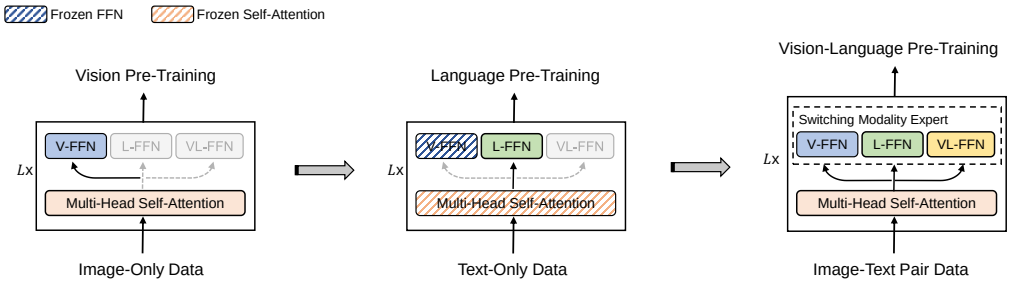


Figure 2: Stagewise pre-training using image-only and text-only corpora. We first pretrain the vision expert (V-FFN) and self-attention module on large-scale image-only data as in BEIT [2]. Then the parameters of vision expert and self-attention module are frozen, and we train the language expert (L-FFN) by masked language modeling on large amounts of text-only data. Finally, we train the whole model with vision-language pre-training.

Image Representations Following vision Transformers [12, 43, 2], the 2D image $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$ is split and reshaped into $N = HW/P^2$ patches $\boldsymbol{v}^p \in \mathbb{R}^{N \times (P^2 C)}$, where C is the number of channels, (H, W) is the resolution of the input image, and (P, P) is the patch resolution. The image patches are then flattened into vectors and are linearly projected to obtain patch embeddings. We also prepend a learnable special token `[I_CLS]` to the sequence. Finally, image input representations are obtained via summing patch embeddings, learnable 1D position embeddings $\boldsymbol{V}_{pos} \in \mathbb{R}^{(N+1) \times D}$ and image type embedding $\boldsymbol{V}_{type} \in \mathbb{R}^D$: $\boldsymbol{H}_0^v = [\boldsymbol{v}_{\texttt{[I_CLS]}}, \boldsymbol{V}\boldsymbol{v}_i^p, \ldots, \boldsymbol{V}\boldsymbol{v}_N^p] + \boldsymbol{V}_{pos} + \boldsymbol{V}_{type}$, where $\boldsymbol{H}_0^v \in \mathbb{R}^{(N+1) \times D}$, linear projection $\boldsymbol{V} \in \mathbb{R}^{(P^2 C) \times D}$.

Text Representations Following BERT [10], we tokenize the text to subword units by WordPiece [46]. A start-of-sequence token (`[T_CLS]`) and a special boundary token (`[T_SEP]`) are added to the text sequence. Text input representations $\boldsymbol{H}_0^w \in \mathbb{R}^{(M+2) \times D}$ are computed via summing the corresponding word embedding, text position embedding and text type embedding $\boldsymbol{H}_0^w = [\boldsymbol{w}_{\texttt{[T_CLS]}}, \boldsymbol{w}_i, \ldots, \boldsymbol{w}_M, \boldsymbol{w}_{\texttt{[T_SEP]}}] + \boldsymbol{T}_{pos} + \boldsymbol{T}_{type}$. M indicates the length of tokenized subword units.

Image-Text Representations We concatenate image and text input vectors to form the image-text input representations $\boldsymbol{H}_0^{vl} = [\boldsymbol{H}_0^w; \boldsymbol{H}_0^v]$

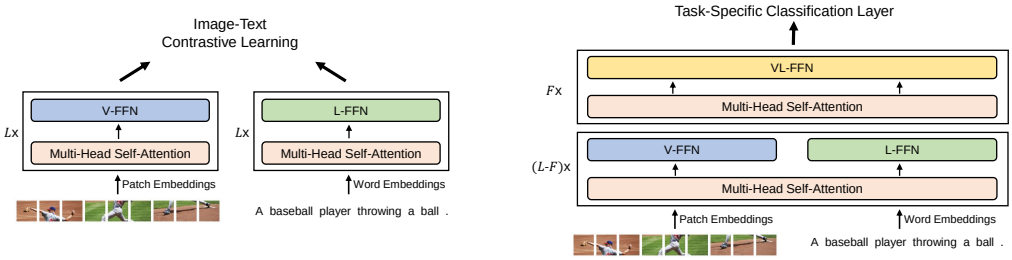
3.2 Mixture-of-Modality-Experts Transformer

Inspired by mixture-of-experts networks [39, 13], we propose a general-purpose multimodal Transformer for vision-language tasks, namely MoME Transformer, to encode different modalities. MoME Transformer introduces mixture of modality experts as a substitute of the feed forward network of standard Transformer. Given previous layer's output vectors $\boldsymbol{H}_{l-1}, l \in [1, L]$, each MoME Transformer block captures modality-specific information by switching to different modality expert, and employs multi-head self-attention (MSA) shared across modalities to align visual and linguistic contents. LN is short for layer normalization.

$$\boldsymbol{H}_l' = \text{MSA}(\text{LN}(\boldsymbol{H}_{l-1})) + \boldsymbol{H}_{l-1} \tag{1}$$

$$\boldsymbol{H}_l = \text{MoME-FFN}(\text{LN}(\boldsymbol{H}_l')) + \boldsymbol{H}_l' \tag{2}$$

MoME-FFN selects an expert among multiple modality experts to process the input according to the modality of the input vectors $\boldsymbol{H}_l'$ and the index of the Transformer layer. Specifically, there are three modality experts: vision expert (V-FFN), language expert (L-FFN) and vision-language expert (VL-FFN). If the input is image-only or text-only vectors, we use vision expert for encoding images and language expert for encoding text. If the input consists of vectors of multiple modalities, such as the vectors of image-text pair, we employ vision expert and language expert to encode the respective modality vectors at the bottom Transformer layers. Vision-language expert is then used at the top layers to capture more modality interaction. Given the three types of input vectors, we obtain image-only, text-only and image-text contextualized representations.



(a) Fine-Tuning on VL Retrieval Tasks

(b) Fine-Tuning on VL Classification Tasks

Figure 3: Fine-tuning VLMo on vision-language retrieval and classification tasks. The model can be fine-tuned as a dual encoder to separately encode image and text for retrieval tasks. VLMo can also be used as a fusion encoder to handle interaction of image-text pairs for classification tasks.

3.3 Pre-Training Tasks

VLMo is jointly pretrained by image-text contrastive learning on the image and text representations, masked language modeling and image-text matching on the image-text pair representations with shared parameters.

Image-Text Contrast Given a batch of N image-text pairs, image-text contrastive learning aims to predict the matched pairs from $N \times N$ possible image-text pairs. There are $N^2 - N$ negative image-text pairs within a training batch.

The final output vectors of `[I_CLS]` token and `[T_CLS]` token are used as the aggregated representation of the image and text, respectively. Followed by a linear projection and normalization, we obtain image vectors $\{\hat{\boldsymbol{h}}_i^v\}_{i=1}^N$ and text vectors $\{\hat{\boldsymbol{h}}_i^w\}_{i=1}^N$ in a training batch to compute image-to-text and text-to-image similarities:

$$s_{i,j}^{i2t} = \hat{\boldsymbol{h}}_i^{v\intercal} \hat{\boldsymbol{h}}_j^w, \; s_{i,j}^{t2i} = \hat{\boldsymbol{h}}_i^{w\intercal} \hat{\boldsymbol{h}}_j^v \tag{3}$$

$$p_i^{i2t} = \frac{\exp(s_{i,i}^{i2t}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{i2t}/\sigma)}, \; p_i^{t2i} = \frac{\exp(s_{i,i}^{t2i}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{t2i}/\sigma)} \tag{4}$$

Where $s_{i,j}^{i2t}$ represents image-to-text similarity of image of i-th pair and text of j-th pair, $s_{i,j}^{t2i}$ is the text-to-image similarity. $\hat{\boldsymbol{h}}_i^w \in \mathbb{R}^D$ and $\hat{\boldsymbol{h}}_j^v \in \mathbb{R}^D$ indicate the normalized vectors of i-th text and j-th image, σ is a learned temperature parameter. p_i^{i2t} and p_i^{t2i} are the $\mathrm{softmax}$-normalized similarities. Cross-entropy losses over image-to-text and text-to-image similarities are used to train the model.

Masked Language Modeling Following BERT [10], we randomly choose tokens in the text sequence, and replace them with the `[MASK]` token. The model is trained to predict these masked tokens from all the other unmasked tokens and vision clues. We use 15% masking probability as in BERT. The final output vectors of masked tokens are fed into a classifier over the whole text vocabulary with cross-entropy loss.

Image-Text Matching Image-text matching aims to predict whether the image and text is matched. We use the final hidden vector of the `[T_CLS]` token to represent the image-text pair, and feed the vector into a classifier with cross-entropy loss for binary classification. Inspired by ALBEF [23], we sample hard negative image-text pairs based on the contrastive image-to-text and text-to-image similarities.

3.4 Stagewise Pre-Training

We introduce a stagewise pre-training strategy, which leverages large-scale image-only and text-only corpus to improve the vision-language model. As present in Figure 2, we first perform vision pre-training on image-only data, and then perform language pre-training on text-only data to learn general

Model	# Pretrain Images	VQA		NLVR2	
		test-dev	test-std	dev	test-P
Base-Size Models Pretrained on COCO, VG, SBU and CC datasets					
UNITER-Base [3]	4M	72.70	72.91	77.18	77.85
VILLA-Base [14]	4M	73.59	73.67	78.39	79.30
UNIMO-Base [25]	4M	73.79	74.02	-	-
ViLT-Base [20]	4M	71.26	-	75.70	76.13
ALBEF-Base [23]	4M	74.54	74.70	80.24	80.50
VLMo-Base	4M	**76.64**	**76.89**	**82.77**	**83.34**
Large-Size Models Pretrained on COCO, VG, SBU and CC datasets					
UNITER-Large [3]	4M	73.82	74.02	79.12	79.98
VILLA-Large [14]	4M	74.69	74.87	79.76	81.47
UNIMO-Large [25]	4M	75.06	75.27	-	-
VLMo-Large	4M	**79.94**	**79.98**	**85.64**	**86.86**
Models Pretrained on More Data					
VinVL-Large [48]	5.7M	76.52	76.60	82.67	83.98
SimVLM-Large [45]	1.8B	79.32	79.56	84.13	84.84
SimVLM-Huge [45]	1.8B	80.03	80.34	84.53	85.15
Florence-Huge [47]	900M	80.16	80.36	-	-
VLMo-Large++	1.0B	**82.88**	**82.78**	**88.62**	**89.54**

Table 1: Fine-tuning results of base-size and large-size VLMo on vision-language classification datasets. VLMo-Large++ is the model trained on one billion noisy image-text pairs with a larger batch size. We report vqa-score on VQA test-dev and test-standard split, and report accuracy for NLVR2 development and public test set (test-P).

image and text representations. The model is used to initialize the vision-language pre-training to learn the alignment of visual and linguistic information. For vision pre-training, we train the attention module and vision expert of MoME Transformer as in BEiT [2] on image-only data. We directly utilize the pretrained parameters of BEiT to initialize the attention module and vision expert. For language pre-training, we freeze parameters of the attention module and vision expert, and utilize masked language modeling [10] to optimize the language expert on text-only data. Compared with image-text pairs, image-only and text-only data are easier to collect. In addition, text data of image-text pairs is usually short and simple. Pre-training on image-only and text-only corpus improves the generalization on complex pairs.

3.5 Fine-Tuning VLMo on Downstream Tasks

As present in Figure 3, our model can be fine-tuned to adapt to various vision-language retrieval and classification tasks.

Vision-Language Classification For classification tasks such as visual question answering and visual reasoning, VLMo is used as a fusion encoder to model modality interaction of images and text. We use the final encoding vector of the token `[T_CLS]` as the representation of the image-text pair, and feed it to a task-specific classifier layer to predict the label.

Vision-Language Retrieval For retrieval tasks, VLMo can be used as a dual encoder to encode images and text separately. During fine-tuning, our model is optimized for the image-text contrastive loss. During inference, we compute representations of all images and text, and then use dot product to obtain image-to-text and text-to-image similarity scores of all possible image-text pairs. Separate encoding enables a much faster inference speed than fusion-encoder-based models.

4 Experiments

We pretrain our model using large-scale image-text pairs and evaluate the model on visual-linguistic classification and retrieval tasks.

Model	# Pretrain Images	MSCOCO (5K test set)						Flickr30K (1K test set)					
		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
Base-Size Models Pretrained on COCO, VG, SBU and CC datasets													
UNITER-Base	4M	64.4	87.4	93.1	50.3	78.5	87.2	85.9	97.1	98.8	72.5	92.4	96.1
VILLA-Base	4M	-	-	-	-	-	-	86.6	97.9	99.2	74.7	92.9	95.8
ViLT-Base	4M	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
ALBEF-Base‡	4M	73.1	91.4	96.0	56.8	81.5	89.2	**94.3**	**99.4**	99.8	**82.8**	**96.7**	**98.4**
VLMO-Base†	4M	**74.8**	**93.1**	**96.9**	**57.2**	**82.6**	**89.8**	92.3	**99.4**	**99.9**	79.3	95.7	97.8
Large-Size Models Pretrained on COCO, VG, SBU and CC datasets													
UNITER-Large	4M	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-Large	4M	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
VLMO-Large†	4M	**78.2**	**94.4**	**97.4**	**60.6**	**84.4**	**91.0**	**95.3**	**99.9**	**100.0**	**84.5**	**97.3**	**98.6**
Models Pretrained on More Data													
VinVL-Large	5.7M	75.4	92.9	96.2	58.8	83.5	90.3	-	-	-	-	-	-
ALIGN-Large†	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
Florence-Huge†	900M	81.8	95.2	-	63.2	85.7	-	**97.2**	99.9	-	87.9	98.1	-
VLMO-Large++†	1.0B	**83.1**	**96.0**	**98.2**	**65.2**	**86.5**	**92.2**	96.8	**100.0**	**100.0**	**88.1**	**98.4**	**99.3**

Table 2: Fine-tuning results of text-retrieval (TR) and image-retrieval (IR) on COCO and Flickr30K. †: ALIGN, Florence and our model encode images and text separately, and then employ a shallow interaction (dot product) to obtain the similarity scores. ‡: ALBEF first encodes images and text separately to obtain the top-k candidates, and then feed these representations into a fusion encoder to rerank the candidates. The others require to encode all image-text combinations by a fusion encoder. VLMO-Large++ represents the model trained on one billion noisy image-text pairs with a larger batch size.

4.1 Pre-Training Setup

Following previous work [3, 20], our pre-training data consists of four image captioning datasets: Conceptual Captions (CC) [38], SBU Captions [32], COCO [27] and Visual Genome (VG) [21] datasets. There are about 4M images and 10M image-text pairs in the pre-training data.

Our models adopt the same network configuration as ViT [12] and BEiT [2]. VLMO-Base consists of 12-layer Transformer blocks with 768 hidden size and 12 attention heads. VLMO-Large is a 24-layer Transformer network with 1024 hidden size and 16 attention heads. VLMO-Base uses vision-language expert on the top two Transformer layers, and VLMO-Large introduces vision-language expert on the top three layers. For images, the input resolution is 224×224 and the patch size is 16×16 during pre-training. We apply RandAugment [9] to the input images. The tokenizer of the uncased version of BERT is employed to tokenize the text. The maximum text sequence length is set to 40. We also employ whole word masking for the masked language modeling pre-training task. We pretrain the models for 200k steps with 1024 batch size. We utilize AdamW [29] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The peak learning is 2e-4 for the base-size model, 5e-5 for the large-size model. Weight decay is set to 0.01. We use linear warmup over the first 2.5k steps and linear decay.

4.2 Training on Larger-scale Datasets

We scale up vision-language representation learning by training VLMO on one billion noisy web image-text pairs with a larger batch size. We first pretrain the model for 200k steps with 16k batch size, and then continue train the model for 100k steps with 32k batch size. The other hyper-parameters are the same as the training on 4M data. Please refer to the supplementary material for more details of hyper-parameters used for pre-training and fine-tuning.

4.3 Evaluation on Vision-Language Classification Tasks

We first conduct fine-tuning experiments on two widely used classification datasets: visual question answering [15] and natural language for visual reasoning [41]. The model is fine-tuned as a fusion encoder to model deeper interaction.

Stagewise Pre-Training	NLVR2		Flickr30k	
	dev	test-P	TR	IR
Image-Only Pre-Training	80.33	81.06	95.60	87.69
Image-Only + Text-Only Pre-Training	**82.09**	**82.49**	**95.67**	**88.52**

Table 3: Ablation studies of stagewise pre-training, i.e., different initialization for vision-language pre-training. We report the average of R@1, R@5 and R@10 for Flickr30k. Results of NLVR2 are averaged over three runs.

	Pre-Training Tasks			Transformer			NLVR2		Flickr30k	
	ITC	ITM	MLM	Std TRM	MoME	MoME−VLExp	dev	test-P	TR	IR
[1]	✓	✗	✗	✗	✓	✗	58.51	58.83	92.23	84.24
[2]	✓	✗	✓	✗	✓	✗	73.91	73.75	94.07	85.82
[3]	✓	✓	✗	✗	✓	✗	76.46	76.19	94.37	85.67
[4]	✓	✓	✓	✓	✗	✗	78.81	79.27	93.37	85.73
[5]	✓	✓	✓	✗	✗	✓	79.58	80.11	94.50	86.69
[6]	✓	✓	✓	✗	✓	✗	**80.13**	**80.31**	**95.17**	**87.25**

Table 4: Ablation studies of MoME Transformer and vision-language pre-training tasks. "ITC" is short for image-text contrastive loss, "ITM" is image-text matching, and "MLM" is masked language modeling. "Std TRM" is short for standard Transformer, and "MoME−VLExp" is MoME without VL experts. The average of R@1, R@5 and R@10 is reported for Flickr30k. Results of NLVR2 are averaged over three runs.

Visual Question Answering (VQA) For VQA, a natural image and a question are given, the task is to generate/choose the correct answer. We train and evaluate the model on VQA 2.0 dataset [15]. Following common practices, we convert VQA 2.0 to a classification task, and choose the answer from a shared set consists of $3,129$ answers. We use the final encoding vector of the `[T_CLS]` token as the representation of the image-question pair and feed it to a classifier layer to predict the answer.

Natural Language for Visual Reasoning (NLVR2) The NLVR2 [41] dataset requires the model to predict whether a text description is true about a pair of images. Following OSCAR [26] and VinVL [48], we convert the triplet input to two image-text pairs, each containing the text description and one image. We concatenate the final output vectors of the `[T_CLS]` token of the two input pairs. The concatenated vector is then fed into a classification layer to predict the label.

We present the results of VL classification tasks in Table 1. VLMo achieves state-of-the-art performance and substantially outperforms previous methods. Our large-size model even outperforms SimVLM-Huge [45] and Florence-Huge [47] by a large margin, which consists of more parameters and are also trained on larger-scale image-text pairs. Our model uses a simple linear projection to embed images as in ViLT [20]. This leads to a significant speedup compared with previous models using image region features, which are extracted by an off-the-shelf object detector [30, 40, 3, 14, 25, 48].

4.4 Evaluation on Vision-Language Retrieval Tasks

The retrieval tasks contain image-to-text retrieval and text-to-image retrieval. We evaluate the model on the widely used COCO [27] and Flickr30K [33] datasets, and use the Karpathy split [19] for both datasets. The model is used as a dual encoder for retrieval tasks. We encode images and text separately and compute their similarity scores by the dot product of image and text vectors.

As present in Table 2, VLMo achieves competitive performance with previous fusion-encoder-based models while having a much faster speed. Fusion-encoder-based models need to jointly encode all possible image-text pairs to compute their similarity scores, which requires quadratic time complexity. Moreover, our large-size model even outperforms the huge-size model of Florence [47], which also trained on massive image-text pairs using a larger batch size. VLMo pre-training can effectively leverage larger-scale noisy pairs and benefit from large batch training.

4.5 Ablation Studies

Stagewise Pre-Training We first conduct ablation experiments of stagewise pre-training. ViLT [20] shows that using the ViT [12] model pretrained on image-only data as the initialization achieves better performance than the BERT model pretrained on text-only data. Therefore we start experiments with image-only pre-training. We compare using image-only pre-training, and image-only pre-training plus text-only pre-training as the initialization. For image-only pre-training, we directly use the parameters of BEiT-Base to initialize the self-attention module and all modality experts. For image-only pre-training plus text-only pre-training, we use pretrained parameters of BEiT-Base to initialize the vision expert and self-attention module of MoME Transformer, and then pretrain its language expert on text corpora. As shown in Table 3, image-only pre-training plus text-only pre-training improves our vision-language model. We also have tried to perform vision-language pre-training with random initialization but obtain a relatively low accuracy on downstream tasks. Stagewise pre-training effectively leverages large-scale image-only and text-only corpus, and improves our vision-language pre-training. Moreover, given the limited size of image-text pairs we used during pre-training, stage-wise pre-training on image-only and text-only data alleviates the need for image-text pair data.

MoME Transformer We also conduct ablation experiments of MoME Transformer. We employ ViT-Base to initialize the models for the ablation experiments. As present in Table 4, using MoME Transformer achieves better performance than standard Transformer for both retrieval and classification tasks. In addition, we also analyse the contribution of vision-language expert (VL-FFN) used in MoME Transformer. We remove the vision-language expert used in the top Transformer layers. Experimental results demonstrate that the introduction of vision-language expert improves the model. Using vision-language expert captures more modality interaction.

Pre-Training Tasks We perform ablation studies to analyse the contribution of different pre-training tasks, and the results are presented in Table 4. Compared with the model trained only using image-text contrastive loss, our unified training performs much better across classification and retrieval tasks. Introducing image-text matching with hard negative mining also greatly improves the model. This demonstrates the effectiveness of our unified-training framework with MoME Transformer. In addition, experimental results show that masked language modeling positively contribute to our model. Please refer to the supplementary material for more ablation studies.

5 Conclusion

In this work, we propose a unified vision-language pretrained model VLMo, which jointly learns a dual encoder and a fusion encoder with a shared MoME Transformer backbone. MoME introduces a pool of modality experts to encode modality-specific information, and aligns different modalities using the shared self-attention module. The unified pre-training with MoME enables the model to be used as a dual encoder for efficient vision-language retrieval, or as a fusion encoder to model cross-modal interactions for classification tasks. We also show that stagewise pre-training that leverages large-scale image-only and text-only corpus greatly improves vision-language pre-training. Experimental results demonstrate that VLMo outperforms previous state-of-the-art models on various vision-language classification and retrieval benchmarks.

In the future, we would like to work on improving VLMo from the following perspectives:

- We will scale up the model size used in VLMo pre-training.
- We are also interested in fine-tuning VLMo for vision-language generation tasks, such as image captioning, following the method proposed in UniLM [11].
- We are going to explore to what extent vision-language pre-training can help each other modality, especially as the shared MoME backbone naturally blends in text and image representations.
- We can extend the proposed model to integrate more modalities (e.g., speech, video, and structured knowledge), supporting general-purpose multimodal pre-training.

References

[1] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR, 2020. URL `http://proceedings.mlr.press/v119/bao20a.html`.

[2] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. URL `https://arxiv.org/abs/2106.08254`.

[3] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. doi: 10.1007/978-3-030-58577-8_7. URL `https://doi.org/10.1007/978-3-030-58577-8_7`.

[4] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xianling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training. *CoRR*, abs/1909.10481, 2019.

[5] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2021.naacl-main.280`.

[6] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. XLM-E: Cross-lingual language model pre-training via ELECTRA. *ArXiv*, abs/2106.16138, 2021.

[7] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf`.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.747`.

[9] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 3008–3017. Computer Vision Foundation / IEEE, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

[11] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054, 2019.

[12] 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. *preprint arXiv:2010.11929*, 2020.

[13] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, abs/2101.03961, 2021. URL `https://arxiv.org/abs/2101.03961`.

[14] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL `https://doi.org/10.1109/CVPR.2017.670`.

[16] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. URL `https://arxiv.org/abs/2004.00849`.

[17] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12976–12985. Computer Vision Foundation / IEEE, 2021.

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. URL `http://proceedings.mlr.press/v139/jia21b.html`.

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015.

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. URL `http://proceedings.mlr.press/v139/kim21k.html`.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[23] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. URL `https://arxiv.org/abs/2107.07651`.

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. URL http://arxiv.org/abs/1908.03557.

[25] 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. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2592–2607. Association for Computational Linguistics, 2021.

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. doi: 10.1007/978-3-030-58577-8_8. URL https://doi.org/10.1007/978-3-030-58577-8_8.

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[28] 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. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

[31] Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736, 2021. URL https://arxiv.org/abs/2106.13736.

[32] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

[33] 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 *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.303. URL https://doi.org/10.1109/ICCV.2015.303.

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://s3-us-west-2.amazonaws.com/openaiassets/research-covers/language-unsupervised/languageunderstandingpaper.pdf.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031. URL https://doi.org/10.1109/TPAMI.2016.2577031.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1238/.

[39] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

[40] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

[41] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1644. URL https://doi.org/10.18653/v1/p19-1644.

[42] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1514. URL https://doi.org/10.18653/v1/D19-1514.

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan,

and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need`.

[45] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021. URL `https://arxiv.org/abs/2108.10904`.

[46] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

[47] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. URL `https://arxiv.org/abs/2111.11432`.

[48] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html`.

[49] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please refer to the conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report average results of multiple runs.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]