Unified Vision-Language Pre-Training for Image Captioning and VQA

Luowei Zhou,[1] **Hamid Palangi,**[2] **Lei Zhang,**[3] **Houdong Hu,**[4] **Jason J. Corso,**[1] **Jianfeng Gao**[2]

[1]University of Michigan, [2]Microsoft Research, [3]Microsoft Cloud & AI, [4]Microsoft AI & Research

{luozhou, jjcorso}@umich.edu, {hpalangi, leizhang, houhu, jfgao}@microsoft.com

Abstract

This paper presents a unified Vision-Language Pre-training (VLP) model. The model is *unified* in that (1) it can be fine-tuned for either vision-language generation (*e.g.*, image captioning) or understanding (*e.g.*, visual question answering) tasks, and (2) it uses a shared multi-layer transformer network for both encoding and decoding, which differs from many existing methods where the encoder and decoder are implemented using separate models. The unified VLP model is pre-trained on a large amount of image-text pairs using the unsupervised learning objectives of two tasks: bidirectional and sequence-to-sequence (seq2seq) masked vision-language prediction. The two tasks differ solely in what context the prediction conditions on. This is controlled by utilizing specific self-attention masks for the shared transformer network. To the best of our knowledge, VLP is the first reported model that achieves state-of-the-art results on both vision-language generation and understanding tasks, as disparate as image captioning and visual question answering, across three challenging benchmark datasets: COCO Captions, Flickr30k Captions, and VQA 2.0. The code and the pre-trained models are available at https://github.com/LuoweiZhou/VLP.

Introduction

Inspired by the recent success of pre-trained language models such as BERT (Devlin et al. 2018) and GPT (Radford et al. 2018; 2019), there is a growing interest in extending these models to learning cross-modal representations like image-text (Lu et al. 2019; Tan and Bansal 2019) and video-text (Sun et al. 2019b; 2019a), for various vision-language tasks such as Visual Question Answering (VQA) and video captioning, where traditionally tedious task-specific feature designs and fine-tuning are required.

Table 1 summarizes some of the recent works on vision-language pre-training where all the models are unexceptionally built upon Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). These models use a two-stage training scheme. The first stage, called pre-training, learns the contextualized vision-language representations by predicting the masked words or image regions based on their intra-modality or cross-modality relationships

Figure 1: We propose a unified encoder-decoder model for general vision-language pre-training. The pre-trained model is then fine-tuned for image captioning and visual question answering. Thanks to our vision-language pre-training, both training speed and overall accuracy have been significantly improved on the downstream tasks compared to random initialization or language-only pre-training. All the results are evaluated on the validation set of the corresponding dataset.

on large amounts of image-text pairs. Then, in the second stage, the pre-trained model is fine-tuned to adapt to a downstream task.

Although significant improvements have been reported

Type	Method	Domain	Architecture	Downstream Tasks
Understanding-based only	LXMERT (Tan and Bansal 2019), ViLBERT (Lu et al. 2019), UNITER (Chen et al. 2019), VisualBERT (Li et al. 2019b), B2T2 (Alberti et al. 2019), Unicoder-VL (Li et al. 2019a),VL-BERT (Su et al. 2019)	Image	Single-stream or two stream Transformer	Visual question answering Visual commonsense reasoning Image retrieval Grounding referring expressions
Generation-based and understanding-based	VideoBERT (Sun et al. 2019b)	Video	Single-stream Transformer+ Masked Transformer (Zhou et al. 2018)	Zero-shot action classification Video captioning
	CBT (Sun et al. 2019a)	Video	Two-stream Transformer encoder+ Transformer decoder	Action anticipation Video captioning
	Our VLP	Image	Single unified encoder-decoder	Visual question answering Image captioning

Table 1: Comparison between our method and other vision-language pre-training works.

on individual downstream tasks using different pre-trained models, it remains challenging to pre-train a *single, unified* model that is universally applicable, via fine-tuning, to a wide range of vision-language tasks as disparate as vision-language generation (*e.g.*, image captioning) and understanding (*e.g.*, VQA). Most existing pre-trained models are either developed only for understanding tasks, as denoted by "understanding-based only" in Tab. 1, or designed as hybrid models that consist of multiple modality-specific encoders and decoders which have to be trained separately in order to support generation tasks. For example, VideoBERT and CBT in Tab. 1 perform pre-training only for the encoder, not for the decoder. This causes a discrepancy between the cross-modal representations learned by the encoder and the representation needed by the decoder for generation, which could hurt the generality of the model. In this paper, we strive to develop a new method of pre-training a unified representation for both encoding and decoding, eliminating the aforementioned discrepancy. In addition, we expect that such a unified representation would also allow more effective cross-task knowledge sharing, reducing the development cost by eliminating the need of pre-training different models for different types of tasks.

To this end, we propose a unified encoder-decoder model, called the Vision-Language Pre-training (VLP) model, which can be fine-tuned for both vision-language generation and understanding tasks. The VLP model uses a shared multi-layer Transformer network (Vaswani et al. 2017) for encoding and decoding, pre-trained on large amounts of image-caption pairs (Sharma et al. 2018), and optimized for two unsupervised vision-language prediction tasks: bidirectional and sequence to sequence (seq2seq) masked language prediction. The two tasks differ solely in what context the prediction conditions on. This is controlled by utilizing specific self-attention masks for the shared Transformer network. In the bidirectional prediction task, the context of the masked caption word to be predicted consists of all the image regions and all the words on its right and left in the caption. In the seq2seq task, the context consists of all the image regions and the words on the left of the to-be-predicted word in the caption.

The proposed VLP has two main advantages in comparison with the BERT-based models in Tab. 1. First, VLP unifies the encoder and decoder and learns a more universal contextualized vision-language representation that can be more easily fine-tuned for vision-language generation and

understanding tasks, as disparate as image captioning and VQA. Second, the unified pre-training procedure leads to a single model architecture for two distinct vision-language prediction tasks, *i.e.*, bidirectional and seq2seq, alleviating the need for multiple pre-training models for different types of tasks without any significant performance loss in task-specific metrics.

We validate VLP in our experiments on both the image captioning and VQA tasks using three challenging benchmarks: COCO Captions (Chen et al. 2015), Flickr30k Captions (Young et al. 2014), and VQA 2.0 dataset (Goyal et al. 2017). We observe that compared to the two cases where we do not use any pre-trained model or use only the pre-trained language model (*i.e.*, BERT), using VLP significantly speed-ups the task-specific fine-tuning and leads to better task-specific models, as shown in Fig. 1. More importantly, without any bells and whistles, our models achieve state-of-the-art results on both tasks across all three datasets.

Related Work

Language Pre-training. Among numerous BERT variants in language pre-training, we review the two methods that are most relevant to our approach, namely Unified LM or UniLM (Dong et al. 2019) and Multi-Task DNN (MT-DNN) (Liu et al. 2019a). UniLM employs a shared Transformer network which is pre-trained on three language modeling objectives: unidirectional, bidirectional, and sequence-to-sequence. Each objective specifies different binary values in the self-attention mask to control what context is available to the language model. MT-DNN combines multi-task training and pre-training by attaching task-specific projection heads to the BERT network. Our work is inspired by these works and tailored for vision-language tasks in particular.

Vision-Language Pre-training. This has become a nascent research area in the vision-language community. Related works include ViLBERT (Lu et al. 2019) and LXMERT (Tan and Bansal 2019), both of which tackle understanding-based tasks only (e.g., VQA and Retrieval) and share the same two-stream BERT framework with a vision-language co-attention module to fuse the information from both modalities. ViLBERT is tested on a variety of downstream tasks including VQA, referring expression, and image-to-text retrieval. LXMERT only focuses on a particular problem space (*i.e.*, VQA and visual reasoning) and the generalization ability further compromises when the datasets from the down-

stream tasks are also exploited in the pre-training stage. The most similar work to ours is VideoBERT (Sun et al. 2019b), which addresses generation-based tasks (*e.g.*, video captioning) and understanding-based tasks (*e.g.*, action classification). However, it separates the visual encoder and the language decoder and performs pre-training only on the encoder, leaving decoder uninitialized. In contrast, we propose a unified model for both encoding and decoding and fully leverage the benefit of pre-training.

Image Captioning & VQA. Most of the recent works on image captioning are built upon (Anderson et al. 2018), where a language model gets clues for sentence generation through dynamically attending on object regions in the image extracted from pre-trained object detectors. Follow-up works further capture the relationships among object regions by using Graph Convolutional Networks (GCNs) (Yao et al. 2018), incorporating language inductive bias (Yang et al. 2019), or enforcing region grounding between image and text (Lu et al. 2018; Zhou et al. 2019). VQA is another prevalent research area in vision and language. Since its initial proposal (Antol et al. 2015), there has been a significant amount of works proposing model architectures to fuse question and image representations (Kim, Jun, and Zhang 2018; Anderson et al. 2018; Gao et al. 2019), new datasets or models to reduce the dataset bias (Zhang et al. 2016; Goyal et al. 2017; Agrawal et al. 2017) and ground the answer in the question (Lewis and Fan 2019). We use our base architecture to perform both image captioning and VQA with minor model structure differences.

Vision-Language Pre-training

We denote the input image as I and the associated/target sentence description (words) as S. We extract a fixed number N of object regions from the image using an off-the-shelf object detector, denoted as $\{r_1, \ldots, r_N\}$ and the corresponding region features as $R = [R_1, \ldots, R_N] \in \mathbb{R}^{d \times N}$, region object labels (probabilities) as $C = [C_1, \ldots, C_N] \in \mathbb{R}^{l \times N}$, and region geometric information as $G = [G_1, \ldots, G_N] \in \mathbb{R}^{o \times N}$, where d is the embedding size, l indicates the number of the object classes of the object detector, and $o = 5$ consists of four values for top left and bottom right corner coordinates of the region bounding box (normalized between 0 and 1) and one value for its relative area (*i.e.*, ratio of the bounding box area to the image area, also between 0 and 1). The words in S are represented as one-hot vectors which are further encoded to word embeddings with embedding size e: $y_t \in \mathbb{R}^e$ where $t \in \{1, 2, \ldots, T\}$ and T indicates the length of the sentence.

Vision-Language Transformer Network

Our vision-language Transformer network, which unifies the Transformer encoder and decoder into a single model, is depicted in Fig. 2 (left). The model input consists of the class-aware region embedding, word embedding and three special tokens. The region embedding is defined as:

$$r_i = W_r R_i + W_p[\text{LayerNorm}(W_c C_i)|\text{LayerNorm}(W_g G_i)] \tag{1}$$

where $[\cdot|\cdot]$ indicates the concatenation on the feature dimension, LayerNorm represents Layer Normalization. The second term mimics the positional embedding in BERT, but adding extra region class information, and W_r, W_p, W_c, W_g are the embedding weights (the bias term and the nonlinearity term are omitted). Note that here we overload the notation of $r_i \in \mathbb{R}^d$ ($i \in \{1, 2, ..., N\}$) to also represent class-aware region embeddings. In addition, we add segment embeddings to r_i as in BERT where all the regions share the same segment embedding where the values depend on the objectives (*i.e.*, seq2seq and bidirectional, see the following section).

The word embeddings are similarly defined as in (Devlin et al. 2018), adding up y_t with positional embeddings and segment embeddings, which is again overloaded as y_t. We define three special tokens [CLS], [SEP], [STOP], where [CLS] indicates the start of the visual input, [SEP] marks the boundary between the visual input and the sentence input, and [STOP] determines the end of the sentence. The [MASK] tokens indicate the masked words which will be explained in the next section.

Pre-training Objectives

In the BERT masked language modeling objective, 15% of the input text tokens are first replaced with either a special [MASK] token, a random token or the original token, at random with chances equal to 80%, 10%, and 10%, respectively. Then, at the model output, the hidden state from the last Transformer block is projected to word likelihoods where the masked tokens are predicted in the form of a classification problem. Through this reconstruction, the model learns the dependencies in the context and forms a language model. We follow the same scheme and consider two specific objectives: the bidirectional objective (bidirectional) as in BERT and the sequence to sequence objective (seq2seq), inspired by (Dong et al. 2019).

As shown in Fig. 2 (right), the only difference between the two objectives lie in the self-attention mask. The mask used for the bidirectional objective allows unrestricted message passing between the visual modality and the language modality while in seq2seq, the to-be-predicted word cannot attend to the words in the future, *i.e.*, it satisfies the auto-regressive property. More formally, we define the input to the first Transformer block as $H^0 = [r_{\text{[CLS]}}, r_1, \ldots, r_N, y_{\text{[SEP]}}, y_1, \ldots, y_T, y_{\text{[STOP]}}] \in \mathbb{R}^{d \times U}$ where $U = N+T+3$, and then the encoding at different levels of Transformer as $H^l = \text{Transformer}(H^{l-1})$, $l \in [1, L]$. We further define a self-attention mask as $M \in \mathbb{R}^{U \times U}$, where

$$M_{jk} = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases} \quad j, k = 1, \ldots, U. \tag{2}$$

For simplicity, we assume a single attention head in the self-attention module. Then, the self-attention output on H^{l-1}
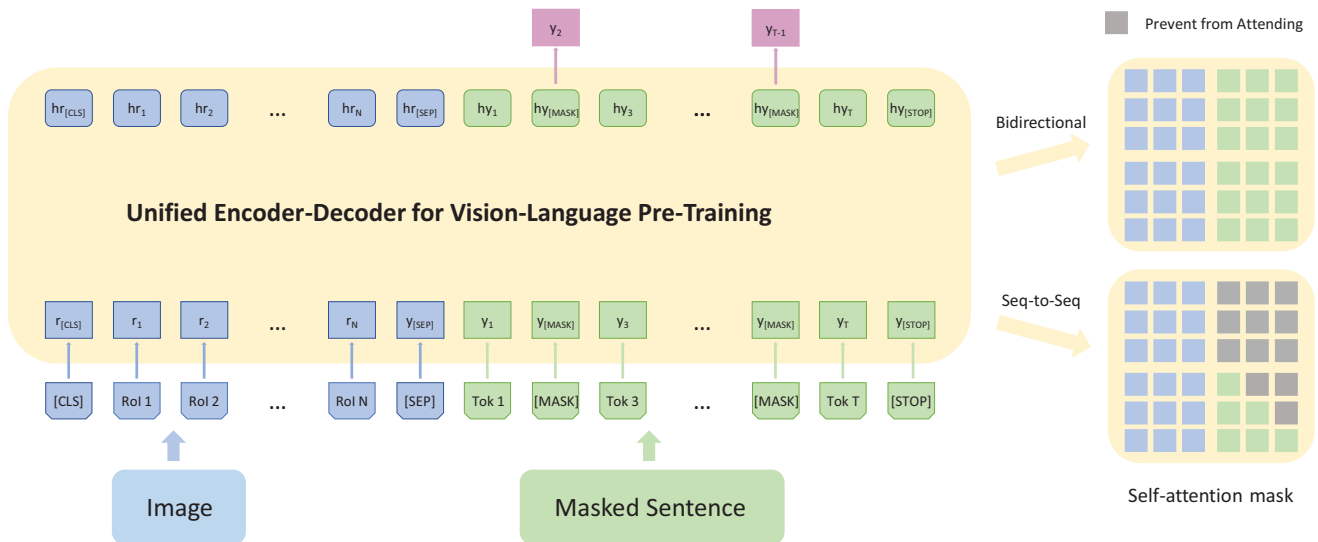
Figure 2: Model architecture for pre-training. The input comprises of image input, sentence input, and three special tokens ([CLS], [SEP], [STOP]). The image is processed as N Region of Interests (RoIs) and region features are extracted according to Eq. 1. The sentence is tokenized and masked with [MASK] tokens for the later masked language modeling task. Our Unified Encoder-Decoder consists of 12 layers of Transformer blocks, each having a masked self-attention layer and feed-forward module, where the self-attention mask controls what input context the prediction conditions on. We implemented two self-attention masks depending on whether the objective is bidirectional or seq2seq. Better viewed in color.

can be formulated as:

$$A^l = \text{softmax}(\frac{Q^\top K}{\sqrt{d}} + M)V^\top, \qquad (3)$$

$$V = W_V^l H^{l-1}, \ Q = W_Q^l H^{l-1}, \ K = W_K^l H^{l-1}, \qquad (4)$$

where W_V^l, W_Q^l, and W_K^l are the embedding weights (the bias terms are omitted). The intermediate variables V, Q, and K indicate values, queries and keys, respectively, as in the self-attention module (Vaswani et al. 2017). A^l is further encoded by a feed-forward layer with a residual connection to form the output H^l. During the pre-training, we alternate per-batch between the two objectives and the proportions of seq2seq and bidirectional are determined by hyper-parameters λ and $1 - \lambda$, respectively.

It is worth noting that in our experiments we find that incorporating the region class probabilities (C_i) into region feature (r_i) leads to better performance than having a masked region classification pretext as in (Lu et al. 2019; Tan and Bansal 2019). Therefore, differing from existing works where masked region prediction tasks are used to refine the visual representation, we indirectly refine the visual representation by utilizing it for masked language reconstruction. We also choose not to use the Next Sentence Prediction task as in BERT, or in our context predicting the correspondence between image and text, because the task is not only weaker than seq2seq or bidirectional but also computationally expensive. This coincidentally agrees with a concurrent work of RoBERTa (Liu et al. 2019b).

Sequence-to-sequence inference. Similar to the way seq2seq training is performed, we can directly apply VLP to sequence-to-sequence inference, in the form of beam search.

More details follow next in the Image Captioning section.

Fine-Tuning for Downstream Tasks

Image Captioning

We fine-tune the pre-trained VLP model on the target dataset using the seq2seq objective. During inference, we first encode the image regions along with the special [CLS] and [SEP] tokens and then start the generation by feeding in a [MASK] token and sampling a word from the word likelihood output (*e.g.*, greedy sampling). Then, the [MASK] token in the previous input sequence is replaced by the sampled word and a new [MASK] token is appended to the input sequence to trigger the next prediction. The generation terminates when the [STOP] token is chosen. Other inference approaches like beam search could apply as well.

Visual Question Answering

We frame VQA as a multi-label classification problem. In this work we focus on open domain VQA where top k most frequent answers are selected as answer vocabulary and used as class labels. Following (Anderson et al. 2018) we set k to 3129.

During the fine-tuning, a multi-layer Perceptron (Linear+ReLU+Linear+Sigmoid) on top of the element-wise product of the last hidden states of [CLS] and [SEP] is learned, similar to (Lu et al. 2019). We optimize the model output scores with respect to the soft answer labels using cross-entropy loss. Note that unlike (Tan and Bansal 2019) where the task-specific objective (*i.e.*, VQA) is exploited during pre-training by using the target datasets (from inten-

Method	COCO				VQA 2.0 (Test-Standard)				Flickr30k			
	B@4	M	C	S	Overall	Yes/No	Number	Other	B@4	M	C	S
BUTD (Anderson et al. 2018)	36.2	27.0	113.5	20.3	65.7	-	-	-	27.3	21.7	56.6	16.0
NBT (with BBox) (Lu et al. 2018)	34.7	27.1	107.2	20.1	-	-	-	-	27.1	21.7	57.5	15.6
GCN-LSTM (spa) (Yao et al. 2018)	**36.5**	27.8	115.6	20.8	-	-	-	-	-	-	-	-
GCN-LSTM (sem)	**36.8**	27.9	116.3	20.9	-	-	-	-	-	-	-	-
GVD (Zhou et al. 2019)	-	-	-	-	-	-	-	-	26.9	22.1	60.1	16.1
GVD (with BBox)	-	-	-	-	-	-	-	-	27.3	22.5	62.3	16.5
BAN (Kim, Jun, and Zhang 2018)	-	-	-	-	70.4	85.8	**53.7**	**60.7**	-	-	-	-
DFAF (Gao et al. 2019)	-	-	-	-	70.3	-	-	-	-	-	-	-
AoANet* (Huang et al. 2019)	37.2	28.4	119.8	21.3	-	-	-	-	-	-	-	-
ViLBERT* (Lu et al. 2019)	-	-	-	-	70.9	-	-	-	-	-	-	-
LXMERT* (Tan and Bansal 2019)	-	-	-	-	72.5	88.2	54.2	63.1	-	-	-	-
Ours												
w/o VLP pre-training (baseline)	35.5	28.2	114.3	21.0	70.0	86.3	52.2	59.9	27.6	20.9	56.8	15.3
seq2seq pre-training only	**36.5**	**28.4**	**117.7**	**21.3**	70.2	86.7	52.7	59.9	**31.1**	**23.0**	**68.5**	**17.2**
bidirectional pre-training only	36.1	28.3	116.5	21.2	**71.3**	**87.6**	53.5	**61.2**	30.5	22.6	63.3	16.9
Unified VLP	**36.5**	**28.4**	116.9	21.2	70.7	87.4	52.1	60.5	30.1	**23.0**	67.4	**17.0**

Table 2: Results on COCO Captions test set (with cross-entropy optimization only, all single models), VQA 2.0 Test-Standard set and Flickr30k test set. *** indicates unpublished works**. B@4 represents for BLEU@4, M for METEOR, C for CIDEr, and S for SPICE. Results on previous works are obtained from the original papers. Top two results on each metric are in bold.

Method	COCO (w/ CIDEr optimization)			
	B@4	M	C	S
BUTD	36.3	27.7	120.1	21.4
GCN-LSTM (spa)	38.2	28.5	127.6	22.0
SGAE (Yang et al. 2019)	38.4	28.4	127.8	22.1
AoANet*	38.9	29.2	129.8	22.4
Ours (Unified VLP)	**39.5**	**29.3**	**129.3**	**23.2**

Table 3: Results on COCO Captions test set (with CIDEr optimization, all single models). *** indicates unpublished works**. Top one result on each metric is in bold.

sive human annotations), our pre-training does not have this requirement and is therefore more general.

Experiments and Results

Data preparation. We conduct pre-training on the Conceptual Captions (CC) dataset (Sharma et al. 2018) which has around 3 million web-accessible images with associated captions. The datasets for downstream tasks include COCO Captions (Chen et al. 2015), VQA 2.0 (Goyal et al. 2017) and Flickr30k (Young et al. 2014). For COCO Captions and Flickr30k, we follow Karpathy's split[1], which gives 113.2k/5k/5k and 29.8k/1k/1k images for train/val/test splits respectively. For VQA 2.0, we split the dataset with the official partition, *i.e.*, 443.8k questions from 82.8k images for training, 214.4k questions from 40.5k images for validation and report the results on Test-Standard set through the official evaluation server. We trim long sentences and pad short sentences to 20 words and all the words are tokenized and numericalized as in BERT (Devlin et al. 2018).

---

[1]cs.stanford.edu/people/karpathy/deepimagesent/caption_datasets.zip

Implementation details. Our Transformer backbone is the same as BERT-base (Devlin et al. 2018). The input of the network consists of image (regions) and the associated/target caption. We represent each input image as 100 object regions extracted from a variant of Faster R-CNN (Ren et al. 2015) pre-trained on Visual Genome (Krishna et al. 2017; Anderson et al. 2018). We take the model output from fc6 layer as the region feature (R_i) and the class likelihood on the 1600 object categories as region object labels (C_i). Note that if not specified, the weights in our BERT model are initialized from UniLM (Dong et al. 2019) pretrained on text corpora only. For caption inference, we use greedy search on the validation set and beam search with beam size 5 on the test set. We perform light model hyperparameter search with the configurations presented in Appendix. λ is set to 0.75 for CC pre-training from light model validation (out of $\{0.25, 0.5, 0.75\}$), and set to 1 for image captioning (*i.e.*, full seq2seq) and 0 for VQA (*i.e.*, full bidirectional).

Model variants and metrics. To demonstrate the effectiveness of our vision-language pre-training, we first include a baseline model without this pre-training. We then include two extreme settings of our model with $\lambda = 1$ (seq2seq pre-training only) and $\lambda = 0$ (bidirectional pre-training only) to study how each objective individually works with different downstream tasks. Our full model conducts joint training on the two objectives. The fine-tuning procedure is performed the same regardless of the pre-training configurations. Regarding evaluation metrics, we use standard language metrics for image captioning, including Bleu@4, METEOR, CIDEr, and SPICE and the official measurement on accuracy for VQA, over different answer types including Yes/No, Number, and Other.

Comparisons against SotAs. Results comparing our methods and SotA methods on the test set are in Tab. 2. We include state-of-the-art published works (upper part of Tab. 2),

Method	COCO				VQA 2.0 (Test-Dev)				Flickr30k			
	B@4	M	C	S	Overall	Yes/No	Number	Other	B@4	M	C	S
From scratch	35.2	27.9	112.5	20.6	67.7	83.5	50.7	58.1	28.4	20.8	53.5	15.2
Init from BERT	34.8	28.1	112.6	20.7	68.6	85.2	50.9	58.3	29.1	21.7	60.4	15.9
Init from UniLM	35.5	28.2	114.3	21.0	69.6	86.1	**52.4**	59.4	27.6	20.9	56.8	15.3
Unified VLP	**36.5**	**28.4**	**116.9**	**21.2**	**70.5**	**87.2**	52.1	**60.3**	**30.1**	**23.0**	**67.4**	**17.0**

Table 4: Impact of different levels of pre-training on downstream tasks. All results are on the test set (Test-Dev for VQA 2.0). Top one result on each metric is in bold.

Method	B@4	M	C	S
From scratch	5.5	9.4	63.8	14.9
Init from BERT	5.7	9.7	66.7	15.3
Init from UniLM	5.8	9.7	67.0	15.5

Table 5: Impact of model weight initializations on pre-training. Results are on Conceptual Captions val set on caption generation.

Method	B@4	M	C	S
Region label as pretext	5.4	9.4	62.2	14.5
Region label probability as input	5.8	9.7	67.0	15.5

Table 6: Comparison between having region class prediction pretext and feeding in class probabilities as a part of the model input. Results are on Conceptual Captions val set.

unpublished works that are currently in submission (middle part), and our methods (lower part). All the image captioning methods are single models, with cross-entropy optimization only for a fair comparison. Our full model (Unified VLP) outperforms SotA methods on three out of four metrics on COCO, overall accuracy on VQA 2.0, and all four metrics on Flickr30k. The improvements are particularly sound on Flickr30k, where we get *5.1% absolute gain* on CIDEr metric and *2.8%* on BLEU@4.

We further perform CIDEr optimization on COCO Captions through Self-Critical Sequence Training (SCST) (Rennie et al. 2017), as in most of the recent image captioning literatures. The results are in Tab. 3 where our full model sets new SotA on all the metrics.

Boost from pre-training. Our full model leads our baseline model by a large margin on most of the metrics thanks to our pre-training. Some noticeable improvements include over *10% absolute gain* on CIDEr metric on Flickr30k, and over *2% gain* on CIDEr on COCO and B@4, METEOR on Flickr30k. Small datasets (*i.e.*, Flickr30k) benefit the most as vision-language pre-training alleviates overfitting issues. Our model variants under the two extreme settings work well as expected on their "favorable" tasks, *i.e.*, seq2seq pre-training alone improves downstream captioning tasks significantly and bidirectional pre-training benefits understanding tasks (*i.e.*, VQA), but not the opposite. They set new SotAs on all metrics except the "Number" accuracy on VQA 2.0. The joint training organically combines the representations learned from the two rather different objectives and yields slightly compromised but decent accuracy on all the downstream tasks. That said, from an engineering perspective, if we can afford having separate pre-training models for generation task or understanding task, we will get the optimal model performance. If we value model architecture and parameter sharing, the joint model is a good trade-off.

Impact of pre-training types. Depending on how the base

model Transformer is initialized, we define four "degrees" of pre-training from weakest to strongest as i) without any pre-training, *i.e.*, base model is trained from scratch, ii) bidirectional language pre-training, *i.e.*, base model is initialized from BERT weights (Devlin et al. 2018), iii) seq2seq and bidirectional language pre-training, *i.e.*, base model is initialized from UniLM weights (Dong et al. 2019) which is our baseline setting, and iv) our full Vision-Language Pre-training. The corresponding fine-tuning results on downstream tasks are presented in Fig. 1 on the val set (full results see Appendix) and Tab. 4 on the test set. As shown from the figure, our vision-language pre-training significantly accelerates the learning process of downstream tasks and contributes to better overall accuracy. It is worth noting that the learning process of VQA is greatly shortened despite that the hidden states associated with tokens [CLS] and [SEP] are not learned during the pre-training. This indicates that the contextualized vision-language representations can generalize to unseen domains and work reasonable well as a warm-start for new tasks.

We also study how the pre-training types 1-3 influence our vision-language pre-training in terms of caption generation. The results on Conceptual Captions val set at epoch 20 are shown in Tab. 5. All the models are trained based on the unified VLP objective ($\lambda = 0.75$) for a fair comparison. We observe that initializing base model with weights transferred from pure language pre-training benefits vision-language pre-training. The training objectives of UniLM are closer to our seq2seq and bidirectional objectives than the ones in BERT and hence we hypothesize that this counts for the slightly larger improvement. Note that our intention here is to demonstrate how different weight initializations can influence pre-training performance rather than pursuing possibly high quantitative scores (with full seq2seq training, CIDEr could climb to 77.2 after training for 30 epochs).

Region object labels as pretext. Existing works (Zhou et al. 2019; Lu et al. 2018) regard region object labels (probabilities) (C_i) as an important auxiliary to enrich image region

	GT sentences		
	GT sentences: People in matching shirts standing under umbrellas in the sun People in the same colorful shirts have umbrellas. A large group of people with an umbrella outside. A group of men standing next to a lot of umbrellas A group of people that are under one umbrella	**Unified VLP** (159.8): A group of people standing under umbrellas in the rain. **Init from UniLM** (59.0): A group of people standing around each other. **Init from BERT** (59.0): A group of people standing around each other.	**Question**: Are they dressed the same? **Correct answer**: Yes **Unified VLP**: Yes **Init from UniLM**: No **Init from BERT**: No
	GT sentences: A man standing in front of a blue wall A man talks on a phone in a room with blue wallpaper A man holding a cell phone standing in front of blue wallpaper with designs and a large wall vent A man on a cell phone by a bright blue wall A man holding a phone to his ear	**Unified VLP** (180.6): A man talking on a cell phone in front of a blue wall. **Init from UniLM** (126.9): A man talking on a cell phone while standing next to a blue wall. **Init from BERT** (59.6): A man talking on a cell phone while wearing a gray shirt.	**Question**: Is the man taking his own picture? **Correct answer**: No **Unified VLP**: No **Init from UniLM**: Yes **Init from BERT**: Yes
	GT sentences: A man standing by a large air gondola that is docked in a station A train is parked as a man at the top of the stairs waits along side it. Small tram bus parked between two stair cases A man standing next to cable car and a flight of stairs A man getting ready to board the trolley car	**Unified VLP** (28.0): A red train is parked in a station. **Init from UniLM** (36.9): A red train with a man standing on the top of it. **Init from BERT** (21.3): A red train car sitting inside of a train station.	**Question**: How many people are here? **Correct answer**: 1 **Unified VLP**: 2 **Init from UniLM**: 2 **Init from BERT**: 2
	GT sentences: Two boaters are white water rafting through rough currents. Two people in a small boat in a body of water There are people on a boat tube in the water Two people riding a raft through some waves Two people in a canoe in some rapids	**Unified VLP** (7.5): A man riding a surfboard on top of a wave. **Init from UniLM** (7.6): A man and a boy are riding a surfboard on a wave. **Init from BERT** (5.4): A man riding a paddle board on top of a wave.	**Question**: What is the person doing? **Correct answer**: kayaking/boating **Unified VLP**: surfing **Init from UniLM**: surfing **Init from BERT**: surfing

Figure 3: Qualitative examples on COCO Captions and VQA 2.0. The first column indicates images from the COCO validation set. The second column shows the five human-annotated ground-truth (GT) captions. The third column indicates captions generated by three of our methods and the corresponding CIDEr scores, where only Unified VLP has vision-language pre-training. The last column shows VQA questions and correct answers associated with the image and answers generated by our models. The top two are successful cases and the bottom two are failed cases. See text for details.

features and here we follow a similar design. We can also instead use these labels for a masked region classification pretext as in (Tan and Bansal 2019). Here we have a comparison over the two design choices. "region label probability as input" is equivalent to our full model Unified VLP and "region label as pretext" is the implementation from (Tan and Bansal 2019). As shown in the results, predicting class labels as a pretext has a negative impact on the pre-training, in terms of captioning performance. We hypothesize that this is because the class labels from the off-the-shelf object detector might be noisy which compromises the learned feature representation. In contrast, our model refines the visual representation through a more reliable masked language modeling and could correct the errors exist in the class labels.

Qualitative results and analyses. Qualitative examples on COCO Captions and VQA 2.0 are shown in Fig. 3. In the first two examples, our full model with vision-language pre-training captures more details in the image, such as "umbrellas" and "a blue wall" than the baseline methods. It also answers questions correctly. In the third example, all the methods dis-identify the gondola as a train due to their visual similarity. When it comes to the question answering, our methods all give correct answers while the GT answer is incorrect (note that there is a person in the gondola). In the fourth example, all the models mistakenly classify the activity as surfing while the correct one is kayaking/boating. This is consistent across both the caption model and the VQA model, which implies that the feature representations are indeed shared across tasks.

Conclusion

This paper presents a unified Vision-Language Pre-training (VLP) model that can be fine-tuned for both vision-language generation and understanding tasks. The model is pre-trained on large amounts of image-text pairs based on two objectives: bidirectional and seq2seq vision-language prediction. The two disparate objectives are fulfilled under the same architecture with parameter sharing, avoiding the necessity of having separate pre-trained models for different types of downstream tasks (*i.e.*, generation-based or understanding-based). In our comprehensive experiments on image captioning and VQA tasks, we demonstrate that the large-scale unsupervised pre-training can significantly speed up the learning on downstream tasks and improve model accuracy. Besides, compared to having separate pre-trained models, our unified model combines the representations

Method	COCO				VQA 2.0				Flickr30k			
	B@4	M	C	S	Overall	Yes/No	Number	Other	B@4	M	C	S
From scratch	34.5	28.1	114.2	21.1	63.4	80.2	46.4	55.2	26.9	20.8	52.1	14.4
Init from BERT	34.6	**28.4**	114.8	21.4	65.1	82.9	48.0	56.1	27.5	21.9	58.4	15.5
Init from UniLM												
w/o VLP pre-training (baseline)	34.5	28.1	113.9	21.3	66.1	83.8	49.7	56.9	27.5	21.5	58.3	15.3
seq2seq pre-training only	**35.3**	**28.4**	**116.7**	**21.5**	66.4	84.6	**50.1**	56.9	28.9	**23.6**	67.0	**17.2**
bidirectional pre-training only	**35.3**	28.3	116.1	21.4	**68.2**	**85.6**	51.9	**59.3**	29.6	23.2	67.2	16.8
Unified VLP	**35.5**	**28.5**	**118.0**	**21.6**	67.4	85.4	**50.1**	58.3	29.7	**23.8**	69.1	**17.6**

Table 7: Results on COCO Captions, VQA 2.0, and Flickr30k validation set. B@4 represents for BLEU@4, M for METEOR, C for CIDEr, and S for SPICE. Top two results on each metric are in bold.

Dataset	Batch Size	Learning Rate	# of Epochs	GPUs	Time per Epoch
CC	64(x8)	1e-4(x8)	30	8x V100	5hr
COCO	64(x8)	3e-5(x8)	30	8x V100	12min
VQA 2.0	64(x2)	2e-5(x2)	20	2x V100	32min
Flickr30k	64(x8)	3e-5(x8)	30	8x V100	3min
COCO (w/o pre-training)	64(x8)	3e-4(x8)	30	8x V100	12min
COCO (SCST training)	16(x4)	1e-6(x4)	30	4x Titan Xp	3hr

Table 8: Model hyper-parameters and training specifications.

learned from different objectives and yields slightly compromised but decent (SotA) accuracy on all the downstream tasks. In our future work, we would like to apply VLP to more downstream tasks, such as text-image grounding and visual dialogue. Methodology-wise, we would want to see how multi-task fine-tuning can be applied to our framework to alleviate interference between different objectives.

Appendix

Results on Downstream Tasks

We include the validation results on fine-tuning tasks in Tab. 7. Note that for VQA 2.0, all the methods here are only trained on the training set while for the results reported on the test set (Tab. 3 and Tab. 4 in the main paper), all the models are trained on both training set and validation set following the practice from early works.

Implementation Details

Region proposal and feature. We use a variant of Faster RCNN model (Ren et al. 2015) with ResNeXt-101 FPN backbone (Xie et al. 2017) for region proposal and feature extraction. The Faster RCNN model is pre-trained on the Visual Genome dataset (Krishna et al. 2017), following the same procedure in (Anderson et al. 2018) for joint object detection (1600 classes) and attribute classification. We set the number of regions per image to exact 100 as suggested in (Jiang et al. 2018). We take the output of the fc6 layer as the feature representation for each region, and fine-tune the fc7 layer.

Model hyper-parameters. The model hyper-parameters on pre-training and fine-tuning are in Tab. 8. The SCST training on COCO is performed after the VLP pre-training and COCO fine-tuning.

Training details. We use the same training optimizer as in BERT (Devlin et al. 2018) and other training hyper-parameters are in Tab. 8. Our VQA models are trained on 2x V100 GPUs, COCO Captions SCST training on 4x Titan Xp GPUs, and all others are on 8x V100 GPUs.

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4971–4980.

Alberti, C.; Ling, J.; Collins, M.; and Reitter, D. 2019. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6639–6648.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 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*, 6904–6913.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. *arXiv preprint arXiv:1908.06954*.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, 1564–1574.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lewis, M., and Fan, A. 2019. Generative question answering: Learning to answer the whole question. In *International Conference on Learning Representations*.

Li, G.; Duan, N.; Fang, Y.; Jiang, D.; and Zhou, M. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019a. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7219–7228.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.

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

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019b. Videobert: A joint model for video and language representation learning.

Tan, H., and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 5987–5995. IEEE.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 684–699.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition*.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8739–8748.

Zhou, L.; Kalantidis, Y.; Chen, X.; Corso, J. J.; and Rohrbach, M. 2019. Grounded video description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.