

# Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training

Gen Li,<sup>1\*</sup> Nan Duan,<sup>2</sup> Yuejian Fang,<sup>1†</sup> Ming Gong,<sup>3</sup> Daxin Jiang<sup>3</sup>

<sup>1</sup>School of Software & Microelectronics, Peking University, Beijing, China

<sup>2</sup>Natural Language Computing, Microsoft Research Asia, Beijing, China

<sup>3</sup>STCA NLP Group, Microsoft, Beijing, China

ligen.li@pku.edu.cn, fangyj@ss.pku.edu.cn, {nanduan, migon, djiang}@microsoft.com

## Abstract

We propose **Unicoder-VL**, a universal encoder that aims to learn joint representations of vision and language in a pre-training manner. Borrow ideas from cross-lingual pre-trained models, such as XLM (Lample and Conneau 2019) and Unicoder (Huang et al. 2019), both visual and linguistic contents are fed into a multi-layer Transformer (Vaswani et al. 2017) for the cross-modal pre-training, where three pre-trained tasks are employed, including Masked Language Modeling (MLM), Masked Object Classification (MOC) and Visual-linguistic Matching (VLM). The first two tasks learn context-aware representations for input tokens based on linguistic and visual contents jointly. The last task tries to predict whether an image and a text describe each other. After pretraining on large-scale image-caption pairs, we transfer Unicoder-VL to caption-based image-text retrieval and visual commonsense reasoning, with just one additional output layer. We achieve state-of-the-art or comparable results on both two tasks and show the powerful ability of the cross-modal pre-training.

## Introduction

In recent years, pre-trained models have made great progress in both computer vision (CV) and natural language processing (NLP) communities.

In CV, pre-trained models, such as VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016), are usually trained based on CNN using ImageNet (Deng et al. 2009), whose training objective is to predict the categorical label of a given image. For downstream tasks, such as image classification, image retrieval (Karpathy and Fei-Fei 2015) (Lee et al. 2018) and object detection (Ren et al. 2015), the resulting models can extract feature representations for input images, which will be further used in following task-specific models.

In NLP, pre-trained models, such as BERT (Devlin et al. 2018), XLNet (Yang et al. 2019) and RoBERTa (Liu et al. 2019), have achieved state-of-the-art performances in many NLP tasks as well, such as sentiment analysis (Socher et al. 2013), natural language inference (Bowman et al. 2015),

and machine reading comprehension (Rajpurkar et al. 2016). Pre-trained with language modeling, such models can learn general knowledge from large-scale corpus first, and then transfer them to downstream tasks with simple fine-tuning layers.

However, these two types of pre-trained models cannot well handle a cross-modal task directly, if its natural language inputs are long sequences (such as questions), rather than short phrases (such as tags). The reason is two-fold. On one hand, as ImageNet covers categorical labels only, the resulting models cannot deal with long sequences. This is why most such tasks, e.g. visual question answering (VQA) (Antol et al. 2015), visual commonsense reasoning (VCR) (Zellers et al. 2019) and image retrieval (Karpathy and Fei-Fei 2015), still need additional fusion layers to model interaction between visual and linguistic contents. On the other hand, existing NLP pre-trained models can handle long natural language sequences very well. But none of them is trained with visual contents directly.

Motivated by these, we propose a **Universal encoder** for Vision and Language, short for **Unicoder-VL**, a universal encoder based on a multi-layer Transformer (Vaswani et al. 2017), which aims to learn joint representations of vision and language (especially for long sequences) in a pre-training manner. Inspired by BERT and some recent cross-lingual pre-trained models, such as XLM (Lample and Conneau 2019) and Unicoder (Huang et al., 2019), a cross-modal pre-training framework is designed to model the relationships between visual and linguistic contents and learn their joint representations. We use large-scale image-caption pairs in Unicoder-VL training, as such annotations are easy to collect from web, with relatively good quality. Three pre-trained tasks are employed, including Masked Language Modeling (MLM), Masked Object Classification (MOC) and Visual-linguistic Matching (VLM). The first two tasks learn context-aware representations for input tokens based on linguistic and visual contents jointly. The last task tries to predict whether an image and a text describe each other.

As the first step along this new pre-training direction, we evaluate Unicoder-VL on image-text retrieval tasks. From experiments we can see that, by adding a simple fine-tuning layer, Unicoder-VL achieves state-of-the-art results on both

\*Work is done during an internship at Microsoft Research Asia.

†Corresponding author.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

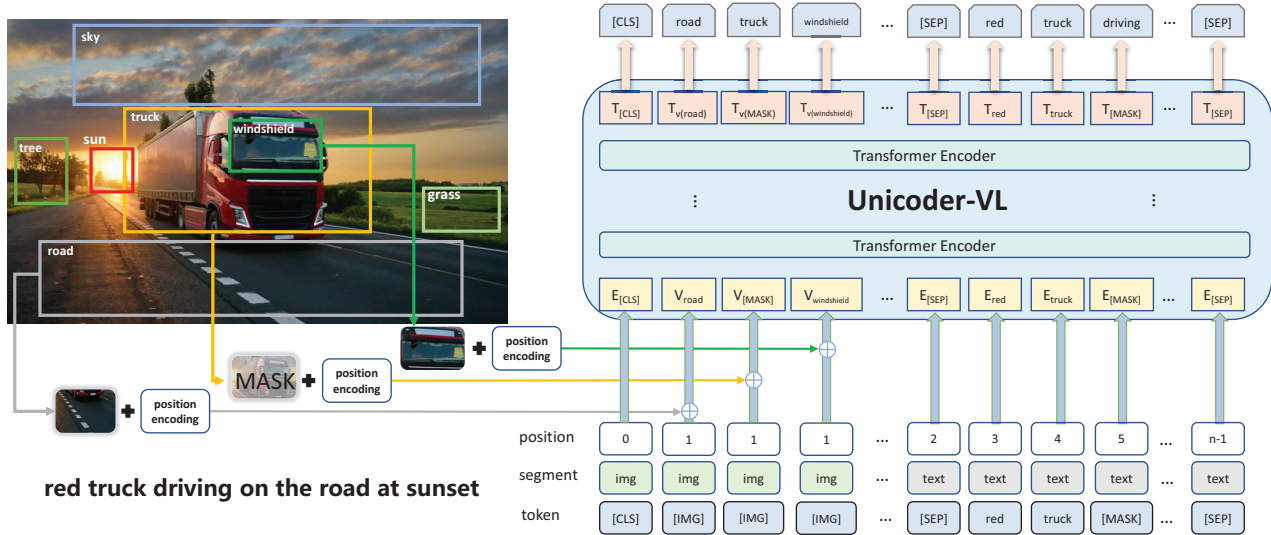


Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn visual and linguistic representation jointly.

MSCOCO (Chen et al. 2015) and Flickr30K (Young et al. 2014), comparing to a bunch of strong baselines. Furthermore, it also shows good performance in a zero-shot setting, which indicates a generalization ability. In VCR, we achieve comparable results with concurrent state-of-the-art works. It shows that cross-modal pre-training improve the ability of visual commonsense reasoning.

The main contributions of our work are summarized as follows. We leverage a multi-layer Transformer to model cross-modal semantic representations. Meanwhile, we propose three well-designed cross-modal pre-training tasks to learn high-level visual representations and capture rich relationships between visual and linguistic contents. We fine-tune our pre-trained model to image-text retrieval and visual commonsense reasoning task and achieve significant improvements, demonstrating the effectiveness of our proposed method. Note, this pre-training method is general and not limited to image-text retrieval tasks. We will move further to evaluate it on more cross-modal tasks, such as image captioning (Anderson et al. 2018), scene graph generation, video classification and video question answering.

## Related Work

### Pre-training for CV Tasks

Most existing pre-trained CV models are based on multi-layer CNN, such as VGG (Simonyan and Zisserman 2014) and ResNet (He et al. 2016), and trained using ImageNet. As ImageNet (Deng et al. 2009) only contains image labels, the resulting pre-trained models cannot deal with cross-modal tasks with long natural language inputs, such as queries in image retrieval and VQA tasks. These tasks pay more attention on visual relations and descriptions rather than what

is the image. By contrast, Unicoder-VL is pre-trained using image-caption pairs. So it is more suitable to these tasks.

### Pre-training for NLP Tasks

Latest pre-trained NLP models are based on multi-layer Transformer, such as GPT (Radford et al. 2018), BERT (Devlin et al. 2018), XLNet (Yang et al., 2019) and RoBERTa(Liu et al. 2019). All of the works are trained using large-scale corpus by language modeling. Such models learn contextualized text representations by predicting word tokens based on their contexts, and can be adapted to downstream tasks by additional fine-tuning.

Since the image is not a sequential data, the autoencoding objective of BERT is very appropriate for visual content. The key question is how to include visual contents in pre-training as well. However, the cross-modal pre-training is not limited to transformer-based models like BERT or XLNet. We leave more exploration in the future.

### Pre-training for Cross-modal Tasks

Very recently, several attempts have been made to pre-train models for cross-modal tasks.

VideoBERT (Sun et al. 2019) is one such method, whose goal is to learn cross-modal representations from videos and their corresponding transcripts. However, instead of using visual features directly in pre-training, it generates a sequence of “visual words” from each video first, and then uses them with transcript words together in LM pre-training. While in Unicoder-VL, we present visual features of objects in the images jointly training with linguistic contents.

Concurrent to our work, several recent released works, such as ViLBERT (Lu et al. 2019), VisualBERT (Li et al. 2019), VL-BERT (Su et al. 2019) and UNITER (Chen et

al. 2019) are pre-training methods on vision-and-language tasks. The concurrent emergency of these research works indicates the importance of deriving a generic pre-trainable representation for cross-modal tasks.

The comparison of these models are:

1) The model of ViLBERT is a two single-modal network applied on input sentences and images respectively, followed by a cross-modal Transformer combining information from the two sources. They propose a co-attentional Transformer layer (Co-TRM) in their model and claim such structure has a better ability to model interactions between visual and linguistic contents. Then the third Transformer fuses them. On the other hand, VisualBERT, Unicoder-VL, VL-BERT and UNITER proposed a single-stream architecture (vanilla BERT structure), which fuses cross-modal information early and freely,

2) a) The masked language model pre-training task is used by all of the above models. b) VisualBERT does not apply the object prediction task. Our model predicts the object labels while the others calculate the KL divergence between the input and output distributions. VL-BERT masked the image before applying by Faster-RCNN. UNITER masks only one modality each time. c) The visual-linguistic matching is used by all of the above models except VL-BERT, which claims this task is of no use.

3) VisualBERT is pre-trained on MSCOCO Captions dataset. ViLBERT, and VL-BERT are all pre-trained on about 3 million Conceptual Captions (Sharma et al. 2018) dataset and then transfer to down-stream tasks. UNITER add about 1 million image-caption pair besides the Conceptual Captions and in-domain MSCOCO Caption and Visual Genome Dense Captions (Krishna et al. 2017) data.

Compare to recent works, we achieve the SOTA results on image-to-text and text-to-image retrieval and VCR, which proves Unicoder-VL’s ability on these tasks.

## Approach

In this section, we first briefly summarize the original BERT model, and present our cross-modal pre-trained model Unicoder-VL, including details of image and text pre-processing and three cross-modal pre-training tasks we used.

### BERT

BERT (Devlin et al. 2018) is a pre-trained model based on multi-layer Transformer (Vaswani et al. 2017). Two tasks are used in pre-training: masked language model and next sentence prediction. In masked language model, BERT tries to predict the identity of each masked word based on all context words. In next sentence prediction, BERT tries to predict whether the second half of the input follows the first half of the input in the corpus, or is a random paragraph. A special token, [CLS], is prepended to every input sequence, and its representation in final layer will be used for the next sentence prediction task.

### Unicoder-VL

The overview of Unicoder-VL is shown in Fig 1. Given a pair of image and sentence, Unicoder-VL takes the visual

regions of the image and textual tokens of the sentence as the input and then encode the input to the linguistic embedding and image embedding. These embeddings are then fed into a multi-layer self-attention Transformer to learn a cross-modality contextualized embedding between visual regions and textual tokens.

**Linguistic Embedding.** Following the text pre-processing of BERT, We tokenize each input text  $\mathbf{w} = \{w_1, \dots, w_T\}$ .  $T$  is the length of the WordPiece (Wu et al. 2016) linguistic input. Besides, as shown in Fig 1, we also add the special tokens [CLS] and [SEP]. For the visual elements, a special [IMG] token is assigned for each one of them. The final representation for each sub-word token is obtained via summing up its word embedding and position embedding, followed by a layer normalization (LN) layer. These embeddings are all initialized from BERT.

**Image Embedding.** For each input image, we first use Faster R-CNN (weights are initialized from (Singh et al. 2018)) to extract the visual features (pooled ROI features) for each region. We also encode the location features with a 5-D vector,  $b = (\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2 - y_1)(x_2 - x_1)}{W \cdot H})$ , where  $(x_1, y_1)$  and  $(x_2, y_2)$  denote the coordinate of the bottom-left and top-right corner and the fraction of image area covered respectively, and  $W, H$  are of the width and height of the input image. Both visual and location features are then fed through a fully-connected (FC) layer, to be projected into the same embedding space. The final visual embedding for each region is obtained by summing up the two FC outputs and then passing through another LN layer. The final image regions are denotes as  $\mathbf{v} = \{v_1, \dots, v_I\}$ .  $I$  is the length of the objects extracted from this image.

We also keep the predicted label of each detected object, which will be used in the object label prediction task. Note that the whole Faster R-CNN model is fixed during training.

**Pre-training Tasks.** We propose three tasks when doing the cross-modal pre-training: Masked Language Modeling (MLM), Masked Object Classification (MOC) and Visual-linguistic Matching (VLM).

**Masked Language Modeling (MLM).** We denote the linguistic input as  $\mathbf{w} = \{w_1, \dots, w_T\}$  and object regions as  $\mathbf{v} = \{v_1, \dots, v_I\}$ , and the mask indices as  $m \in \mathbb{N}^M$ . In MLM, we randomly mask out the input words with probability of 15%, and replace the masked ones  $\mathbf{w}_m$  with special token [MASK]. The goal is to predict these masked words based on the observation of their surrounding words  $\mathbf{w}_{\setminus m}$  and all image regions  $\mathbf{v}$ , by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLM}}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \log P_{\theta}(\mathbf{w}_m | \mathbf{w}_{\setminus m}, \mathbf{v}) \quad (1)$$

where  $\theta$  is the trainable parameters. Each pair  $(\mathbf{w}, \mathbf{v})$  is sampled from the whole training set  $D$ .

**Masked Object Classification (MOC).** Similar to MLM, we also sample image regions and mask their visual features with a probability of 15%. We replace the object feature vector with a zero-initialized vector  $\mathbf{v}_m$  90% of the time, and keep the object feature unchanged in the left 10% time. We simply take the object category with the highest confidence score predicted by the same detection model as the ground-truth label. We first feed the Transformer output of the masked region  $\mathbf{v}_m^{(i)}$  into an FC layer to predict the

scores of  $K$  object classes, which further goes through a soft-max function to be transformed into a normalized distribution  $g_\theta(\mathbf{v}_m^{(i)})$ . The final objective is:

$$\mathcal{L}_{\text{MOC}}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \sum_{i=1}^M \text{CE}(c(\mathbf{v}_m^{(i)}), g_\theta(\mathbf{v}_m^{(i)})) \quad (2)$$

where  $c(\mathbf{v}_m^{(i)}) \in \mathbb{R}^K$  is the one-hot vector of the ground-truth label.

**Visual-linguistic Matching (VLM).** we also learn an instance-level alignment (rather than token/region-level) between the whole image and the sentence via VLM. We take final hidden state of [CLS] to predict whether the linguistic sentence is semantically matched with the visual content, with an additional FC layer. The scoring function is denoted as  $s_\theta(\mathbf{w}, \mathbf{v})$ . During training, we sample both positive and negative image-sentence pairs and learn their matching scores (including negative image and negative sentence). We denote the label as  $y \in \{0, 1\}$ , indicating if the sampled pair is a match. Then

$$\mathcal{L}_{\text{VLM}}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} [y \log s_\theta(\mathbf{w}, \mathbf{v}) + (1 - y) \log(1 - s_\theta(\mathbf{w}, \mathbf{v}))] \quad (3)$$

Overall, we have three training regimes corresponding to the image-text inputs. Our final pre-training objective is the sum of the losses above:

$$\mathcal{L} = (\mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{MOC}}) \cdot I[y = 1] + \mathcal{L}_{\text{VLM}} \quad (4)$$

where  $I[y = 1]$  is an indicator for the label 1 being correct for the image and caption pair.

## Experiments

In this section, we describe how we pre-train our model and show the evaluation details on image-text retrieval task to which we transfer the pre-trained model.

### Pre-training Unicoder-VL

Conceptual Captions dataset (Sharma et al. 2018) contains about 3.3M image and caption pairs harvested from the web, which are very suitable for our cross-modal pre-training. Due to some broken urls, the size of image-caption pairs of Conceptual Captions dataset is about 3M.

Similar to Conceptual Captions, SBU Captions (Ordonez, Kulkarni, and Berg 2011) dataset is also automatically collected from Web and contains 1M image-caption pairs. Due to some broken urls, the size of image-caption pairs of SBU dataset is about 0.8M.

Finally, we use 3.8M image-caption pairs to do pre-training.

Our model has 12 layers of Transformer blocks, where each block has 768 hidden units and 12 self-attention heads. The maximum sequence length is set as 144. We sample 1 negative image or 1 negative caption and then judge whether this image and caption is matching when do the VLM task. The parameters are initialized from BERT-base, which is pre-trained on text data only.

For the visual part, we use fixed 100 RoIs with detection scores higher than 0.2 are selected for each image. If eligible RoIs are less than 100, we simply select the top-100 RoIs, regardless of the detection score threshold.

During Pre-training, our experiments are running on 4 NVIDIA Tesla V100 GPU. Our best performing model is pre-trained for 10 epochs with three training tasks introduced above, using the ADAM optimizer with learning rate of  $1e-4$  with a batch size of 192 with gradient accumulation (every 4 steps). The model will warmup the first 10% of all training steps. We use float16 operations to speed up training and to reduce the memory usage of our models.

### Fine-tune on Downstream Tasks

The pre-trained Unicoder-VL model can be transferred to multiple downstream visual-linguistic tasks, with simple modifications on the input format, output prediction, loss function and training strategy.

**Image-Text Retrieval.** Image-text retrieval is the task of identifying an image from candidates given a caption describing its content, or vice versa. We use two datasets as follows. 1) MSCOCO consists of 123,287 images, and each image contains roughly five textual descriptions. It is split into 82,783 training images, 5,000 validation images and 5,000 testing images. We follow the data split in (Faghri et al. 2017) to add 30,504 images that were originally in the validation set of MSCOCO. 2) Flickr30K contains 31,783 images collected from the Flickr website. Following (Karpathy and Fei-Fei 2015), we split the dataset into 29,783 training images, 1,000 validation images and 1,000 testing images. Besides, we use three evaluation metrics, i.e., R@K (K=1,5,10). R@K is the percentage of ground-truth matchings appearing in the top K-ranked results.

During fine-tuning on image-text retrieval, we formulate it as a ranking problem. we sample 3 negative cases in each matching tasks. Inputs of fine-tuning share the same data preprocessing procedures with pre-training, except that we do not mask word and object in the fine-tuning stage. Similar to the VLM task, we also denote the score function as  $s_\theta(\mathbf{w}, \mathbf{v})$ . We omit this trainable parameter  $\theta$  below. We propose two image-text matching tasks: image-to-text, text-to-image. We use triplet loss and maximize the margin of positive and negative samples after generating the similarity score between two input modalities.

In this study, we focus on the hardest negatives in every sampled examples, following (Faghri et al. 2017). For a positive pair  $(\mathbf{w}, \mathbf{v})$ , the hardest negatives are given by  $\mathbf{v}_h^- = \arg \max_{\mathbf{v}_i \neq \mathbf{v}} s(\mathbf{w}, \mathbf{v}_i)$  and  $\mathbf{w}_h^- = \arg \max_{\mathbf{w}_i \neq \mathbf{w}} s(\mathbf{w}_i, \mathbf{v})$ . So the hardest triplet loss function is:

$$\mathcal{L}_{hard}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{y}^- \in \mathcal{N}_y} \{\max[0, \gamma - s(\mathbf{x}, \mathbf{y}) + s(\mathbf{x}, \mathbf{y}_h^-)]\} \quad (5)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are encodings of two modality,  $\mathcal{N}_y$  is the set of negative samples of  $\mathbf{y}$ .

Finally, we merge these ranking constraints into one loss function:

$$\mathcal{L} = \lambda_1 \sum_{\mathbf{w}, \mathbf{v}} \mathcal{L}_{hard}(\mathbf{w}, \mathbf{v}) + \lambda_2 \sum_{\mathbf{w}, \mathbf{v}} \mathcal{L}_{hard}(\mathbf{v}, \mathbf{w}) \quad (6)$$



Methods	MSCOCO						Flickr30k					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
1K Test set												
DVSA (Karpathy and Fei-Fei 2015)	38.4	69.9	80.5	27.4	60.2	74.8	22.2	48.2	61.4	15.2	37.7	50.5
m-CNN (Ma et al. 2015)	42.8	73.1	84.1	32.6	68.6	82.8	33.6	64.1	74.9	26.2	56.3	69.6
DSPE (Wang, Li, and Lazebnik 2016)	50.1	79.7	89.2	39.6	75.2	86.9	40.3	68.9	79.9	29.7	60.1	72.1
VSE++ (Faghri et al. 2017)	64.7	-	95.9	52.0	-	92.0	52.9	79.1	87.2	39.6	69.6	79.5
SCAN (Lee et al. 2018)	72.7	94.8	98.4	58.8	88.4	94.8	67.4	90.3	95.8	48.6	77.7	85.2
SCG (Shi et al. 2019)	76.6	96.3	99.2	61.4	88.9	95.1	71.8	90.8	94.8	49.3	76.4	85.6
PFAN (Wang et al. 2019)	76.5	96.3	99.0	61.6	89.6	95.2	70.0	91.8	95.0	50.4	78.7	86.1
ViLBERT (Lu et al. 2019) <sup>†</sup>	-	-	-	-	-	-	-	-	-	58.2	84.9	91.5
UNITER (Chen et al. 2019) <sup>†</sup>	-	-	-	-	-	-	84.7	<b>97.1</b>	99.0	71.5	<b>91.2</b>	<b>95.2</b>
Unicoder-VL (zero-shot)	54.4	82.8	90.6	43.4	76.0	87.0	64.3	85.8	92.3	48.4	76.0	85.2
Unicoder-VL (w/o pre-training)	75.1	94.3	97.8	63.9	91.6	96.5	73.0	89.0	94.1	57.8	82.2	88.9
Unicoder-VL	<b>84.3</b>	<b>97.3</b>	<b>99.3</b>	<b>69.7</b>	<b>93.5</b>	<b>97.2</b>	<b>86.2</b>	96.3	<b>99.0</b>	<b>71.5</b>	90.9	94.9
5K Test set												
SCAN (Lee et al. 2018)	50.4	82.2	90.0	38.6	69.3	80.4	-	-	-	-	-	-
SCG (Shi et al. 2019)	56.6	84.5	92.0	39.2	68.0	81.3	-	-	-	-	-	-
UNITER (Chen et al. 2019) <sup>†</sup>	<b>63.3</b>	87.0	<b>93.1</b>	<b>48.4</b>	<b>76.7</b>	<b>85.9</b>	-	-	-	-	-	-
Unicoder-VL	62.3	<b>87.1</b>	92.8	46.7	76.0	85.3	-	-	-	-	-	-

Table 1: Evaluation results on MSCOCO and Flickr30k test set. <sup>†</sup> means the concurrent work.

Followed We use  $\gamma = 0.2$ ,  $\lambda_1 = 1.0$ ,  $\lambda_2 = 1.0$  as the hyper-parameters of loss function. The optimizer is Adam and learning rate is set as  $5e-5$ . The batch size is 192 with gradient accumulation (every 4 steps). We also use float16 operations to speed up training and to reduce the memory usage of our models.

**Zero-shot Image-Text Retrieval.** The previous tasks are all transferring tasks that include dataset specific fine-tuning. In this zero-shot task, we directly apply the pretrained the multi-modal alignment prediction mechanism to image-text retrieval without finetuning. The goal of this task is to demonstrate that the pretraining has developed the ability to ground text and that this can generalize to visual and linguistic variation without any task specific fine-tuning. We directly use the pre-trained Unicoder-VL model and the same alignment prediction objective as a scoring function and test on the same split as the image-text retrieval task described above.

**Visual Commonsense Reasoning.** Given an image, the VCR task presents two problems – visual question answering ( $Q \rightarrow A$ ) and answer justification ( $QA \rightarrow R$ ) – both being posed as multiple choice problems. The holistic setting ( $Q \rightarrow AR$ ) requires both the chosen answer and then the chosen rationale to be correct. The Visual Commonsense Reasoning (VCR) dataset consists of 290k multiple choice QA problems derived from 110k movie scenes. Different from the VQA dataset, VCR integrates object tags into the language providing direct grounding supervision and explicitly excludes referring expressions.

To finetune on this VCR, we concatenate the question and each possible response with semicolons to form four different linguistic inputs and pass each through the model along

with the image.  $\mathbf{w} = \{q_1, \dots, q_n, ;, a_1, \dots, a_n\}$  for  $Q \rightarrow A$  and  $\mathbf{w} = \{q_1, \dots, q_n, ;, a_1^*, \dots, a_n^*, ;, r_1, \dots, r_n\}$ . Here,  $q_0, \dots$  are all question tokens,  $a_0, \dots$  are answer tokens,  $a_0^*$  are answer tokens for the correct answer, and  $r_0$  are rationale tokens.

VCR provides ground truth boxes. For each ground truth box, we select the visual feature with highest intersection over union (IoU) from 100 boxes we extract as the new features. Then we add other visual features left after the features with ground truth boxes until the number is 100.

Since some of objects are referenced in  $Q, A, R$ , we add visual feature  $\mathbf{v}_i$  to these tokens additionally.  $i$  is the object index referenced by the linguistic word.

We also add a projection layer to calculate the score for each pair and the final prediction is a softmax over these four scores. The model is trained under a cross-entropy loss. We trained over 20 epochs with a batch size of 48 and initial learning rate of  $3e-5$ .

## Results and Analysis

### Evaluation Results

**Results on Image-Text Retrieval.** We compare Unicoder-VL with state-of-the-art methods on image retrieval and sentence retrieval tasks in three different settings:

- **zero-shot**, where Unicoder-VL is applied to test set directly, without fine-tuning;
- **task-specific train**, where Unicoder-VL is trained on task-specific training data directly, without pre-training;
- **pre-train + fine-tune**, where Unicoder-VL is further fine-tuned on specific tasks.

Experimental results of both datasets are shown in Tab 1. From Tab 1, the results of the zero-shot setting show that

Unicoder-VL can learn general cross-modal knowledge, which take effects in image retrieval and sentence retrieval tasks directly, without any task-specific fine-tuning. Because the difference between automatically collected Conceptual Captions and human-annotated MSCOCO/Flickr30k, this zero-shot result is lower than the finetuned result. Usually finetuning will help the pre-trained model adapt to a little different downstream dataset.

The results of the task-specific train setting show that Unicoder-VL trained on task-specific training data without pre-training still perform better than most previous approaches. It demonstrates the effectiveness of the self-attention mechanism itself on the image-text retrieval tasks.

The results of the pre-train + fine-tune setting show that this setting can significantly outperform all baselines on all evaluation metrics, which proves the superiority of our cross-modal pre-training method.

Taking R@1 for example, our best result on MSCOCO 1K test set obtains 7.8% and 8.1% absolute improvements against the PFAN approach on sentence retrieval task and image retrieval task, respectively. For MSCOCO 5K test set, we can also significantly outperform all baselines on these two tasks. On the Flickr30k testing set, the experiments show similar achievement. Unicoder-VL achieves new state-of-the-art performance and yield a result of 86.2% and 71.5% on R@1 for sentence retrieval and image retrieval, respectively. Compared with PFAN, we achieve absolute boost of 16.2% on R@1 for sentence retrieval and 21.1% on R@1 for image retrieval. The higher improvement on Flickr30k proves that low-resource task can be improved better with pre-training.

We also compare Unicoder-VL with ViLBERT (Lu et al. 2019) and UNITER(Chen et al. 2019)in the image retrieval and sentence retrieval setting. 10.1 points improvements than ViLBERT show the superiority of Unicoder-VL. UNITER uses 1.8M more image-caption pairs than Unicoder-VL, including in-domain dataset like Visual Genome Caption dataset during pre-training, which may greatly boost the performance of the image-text retrieval. Our Unicoder-VL can still achieve comparable results.

Methods	(Q→A)		(QA→R)		(Q→AR)	
	val	test	val	test	val	test
R2C (Zellers et al. 2019)	63.8	65.1	67.2	67.3	43.1	44.0
VisualBERT (Li et al. 2019) <sup>†</sup>	70.8	71.6	73.2	73.2	52.2	52.4
ViLBERT (Lu et al. 2019) <sup>†</sup>	72.4	73.3	74.5	74.6	54.0	54.8
B2T2 (Alberti et al. 2019) <sup>†</sup>	71.9	72.6	76.0	75.7	54.9	55.0
VL-BERT (Su et al. 2019) <sup>†</sup>	73.7	74.0	74.5	74.8	55.0	55.5
UNITER* (Chen et al. 2019) <sup>†</sup>	72.8	-	75.3	-	54.9	-
Unicoder-VL (w/o pre-training)	71.6	-	73.1	-	52.3	-
Unicoder-VL	72.6	73.4	74.5	74.4	54.5	54.9

Table 2: Results compared to the state-of-the-art methods with single model on VCR dataset by the time of submission. <sup>†</sup> means concurrent works. \* means that the UNITER’s one-stage pre-training result, which is similar to the concurrent work’s setting.

**Results on Visual Commonsense Reasoning (VCR)** Our final results on the VCR task are shown in Tab 2. Pre-training Unicoder-VL only slightly improves the performance. This might be because the pre-training task of image captioning is at the perceptual level, while the VCR task is at the cognitive understanding level. There is a gap between these two data types. Compared with baseline, R2C, we do not use task-specific modules. Instead, we simply add a simple classification layer to Unicoder-VL and jointly train the whole model end-to-end. Unicoder-VL outperforms R2C by large margins, indicating the power of our simple cross-modal architecture. The results without pre-training are slightly lower than results of pre-trained Unicoder-VL. It proves that VCR benefits from cross-modal pre-training. However, due to the difference of VCR dataset and caption dataset, the pre-training will not help too much.

Compared with other concurrent works, i.e. ViLBERT, VisualBERT, B2T2 and VLBERT, our Unicoder-VL achieves comparable performance with state-of-the-art results. It proves that pre-train the transformer-based model with large-scale dataset will yield improvement than previous task-specific methods on visual commonsense reasoning tasks. Note that UNITER proposes a two-stage pre-training for VCR. For a fair comparison, we select the one-stage pre-training result of UNITER here, and it shows similar performance with concurrent works. But two-stage pre-training may be helpful on some very different datasets, like VCR and the caption dataset.

## Discussion

For the pre-training tasks. Unlike VideoBERT (Sun et al. 2019), we do not use image-only inputs since the model fails to converge. But the visual inputs of VideoBERT is actually generated visual words and its objective is still LM pre-training. We assume the true visual inputs without the guidance of linguistic data will damage the pretrained weights of BERT, which is pre-trained on linguistic data only. For future works, we are curious about how we could extend Unicoder-VL to image-only tasks like image-caption, scene graph generation or visual saliency detection.

For image-text retrieval task, the results of Unicoder-VL outperform all the methods without jointly pre-training (actually visual features from ResNet and linguistic word embeddings are pre-trained separately). It demonstrates that this transferring learning can also achieve great performance in cross-modal tasks. However, for image RoI based methods like SCAN(Lee et al. 2018), Unicoder-VL and ViLBERT (Lu et al. 2019), the backbone of Faster-RCNN is still not fine-tuned with the whole model during cross-modal training. We have no idea that whether the performance is better or not if the backbone of detection model is fine-tuned with the cross-modal training and how to do so. We would like to explore these in the future.

We notice that the zero-shot image-text retrieval result of UNITER(Chen et al. 2019) is much higher than ours. The reason is that UNITER uses in-domain dataset including MSCOCO Caption and Visual Genome Caption dataset to pretrain. These datasets are very similar to Flickr30k and it may be not a zero-shot testing. We believe that it

is inappropriate to use in-domain dataset as pre-training dataset unless as the second-stage pre-training dataset because this in-domain dataset is human-annotated (of high quality) but Conceptual Captions and SBU Captions are automatically collected (sometimes not human-like or not related). However, we agree that the performance on these downstream tasks should be enhanced with more high-quality pre-training data.

## Ablation Studies

In this section, we perform ablation experiments in order to better understand the effect of the model size and the pre-train dataset size.

**Effect of Model Size.** We compare the results of Unicoder-VL models when varying Transformer encoder layers. We test our model with 6-layer, 12-layer and 24-layer Transformer encoders. If the number of the layers are less than 12, we simply load the first several layers of pre-trained weights from BERT. As shown in Tab 3, we find that the image-text retrieval tasks benefit from larger models.

Methods	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL (6-layer)	72.4	93.1	96.3	58.1	83.4	90.2
Unicoder-VL (12-layer)	86.2	96.3	99.0	71.5	90.9	94.9
Unicoder-VL (24-layer)	86.5	97.6	99.3	73.6	92.3	95.8

Table 3: Ablation study of the depth of Unicoder-VL with respect to the number of Transformer encoder layers. All of these experiments are fine-tuning on Flickr30k with pre-trained Unicoder-VL.

**Effect of Training Sets Size** We also studied the impact of the size of the pretraining dataset. For this experiment, we take 75% from the full dataset, and pretrain and fine-tune Unicoder-VL using the same setup as above. We can see that the accuracy grows monotonically as the amount of data increases, which suggests that Unicoder-VL may benefit from even more pretraining data. The same experiment results can be observed in ViLBERT (Lu et al. 2019) and UNITER (Chen et al. 2019).

Methods	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL (0)	73.0	89.0	94.1	57.8	82.2	88.9
Unicoder-VL (3M)	82.3	95.1	97.8	68.3	90.3	94.6
Unicoder-VL (4M)	86.2	96.3	99.0	71.5	90.9	94.9

Table 4: Ablation study of the Flickr30k retrieval results of Unicoder-VL with respect to the pre-training dataset size. The number in parentheses is how many image-text pairs we used in pre-training. 0 means without pre-training.

## Conclusion

In this work, we proposed Unicoder-VL for cross-modal tasks. We utilize large-scale image-caption pairs to pre-train Unicoder-VL. We introduce three different pre-training tasks to align the visual and linguistic modalities and learn

better cross-modal representations. When fine-tuning on image and sentence retrieval tasks, our experiment results on Flickr30K and MSCOCO datasets demonstrate that our pre-trained Transformer model can boost retrieval performance significantly. The zero-shots experiments exhibit that Unicoder-VL can learn general cross-modal knowledge, which take effects in image retrieval and sentence retrieval tasks directly, without any task-specific fine-tuning. The VCR experiment shows that cross-modal pre-training improve the ability of visual commonsense reasoning. This pre-training method is general and not limited to these tasks. We do not see any reason preventing it from finding broader cross-modal applications, including video related tasks. Meanwhile, we still have interest on how Unicoder-VL learn from image-only inputs. We will try to extend to some image-only tasks like image-caption and scene graph generation in the future work.

## Acknowledgments

We thank the anonymous reviewers for their helpful comments and discussions. This research is supported by National Natural Science Foundation of China under Grant NO.61672062, NO.61232005.

## References

- 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.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- 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*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- 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*.



- Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612* 2(7):8.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Huang, H.; Liang, Y.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; and Zhou, M. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.
- Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.
- 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.
- Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–216.
- Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vlb: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.
- Ma, L.; Lu, Z.; Shang, L.; and Li, H. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, 2623–2631.
- Ordóñez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, 1143–1151.
- 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](https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf).
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- 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.
- 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.
- Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge aware semantic concept expansion for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5182–5189. International Joint Conferences on Artificial Intelligence Organization.
- Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Singh, A.; Natarajan, V.; Jiang, Y.; Chen, X.; Shah, M.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia—a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.
- Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.
- Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.
- 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.
- Wang, Y.; Yang, H.; Qian, X.; Ma, L.; Lu, J.; Li, B.; and Fan, X. 2019. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 3792–3798. International Joint Conferences on Artificial Intelligence Organization.
- Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5005–5013.
- Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.
- Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New simi-



larity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67–78.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6720–6731.