Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

Gen Li [†*], **Nan Duan** [§], **Yuejian Fang** [†], **Daxin Jiang** [‡], **Ming Zhou** [§]
[†] School of Software & Microelectronics, Peking University, Beijing, China
[‡] STCA NLP Group, Microsoft, Beijing, China
[§] Microsoft Research Asia, Beijing, China
ligen.li@pku.edu.cn, fangyj@ss.pku.edu.cn
{nanduan, djiang, mingzhou}@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 for the cross-modal pre-training, where three pre-trained tasks are employed, including masked language model, masked object label prediction and visual-linguistic matching. 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 amounts of image-caption pairs, we transfer Unicoder-VL to image-text retrieval tasks with just one additional output layer, and achieve state-of-the-art performances on both MSCOCO and Flicker30K.

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 ELMo (Peters et al. 2018), GPT (Radford et al. 2018), BERT (Devlin et al. 2018) and XLNet (Yang 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. VQA (Antol et al. 2015) 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.

Motivated by these, we propose a **Uni**versal en**coder** for **V**ision and **L**anguage, short for **Unicoder-VL**, a universal encoder based on a multi-layer Transformer, 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 (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-XL training, as such annotations are easy to collect from web, with relatively good quality. Three pre-trained tasks are employed, including masked language model, masked object label prediction and visual-linguistic matching. 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 MSCOCO and Flicker30K, comparing to a bunch of strong baselines. Furthermore, it also shows good performance in a zero-shot setting, which indicates a generalization ability.

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 rela-

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

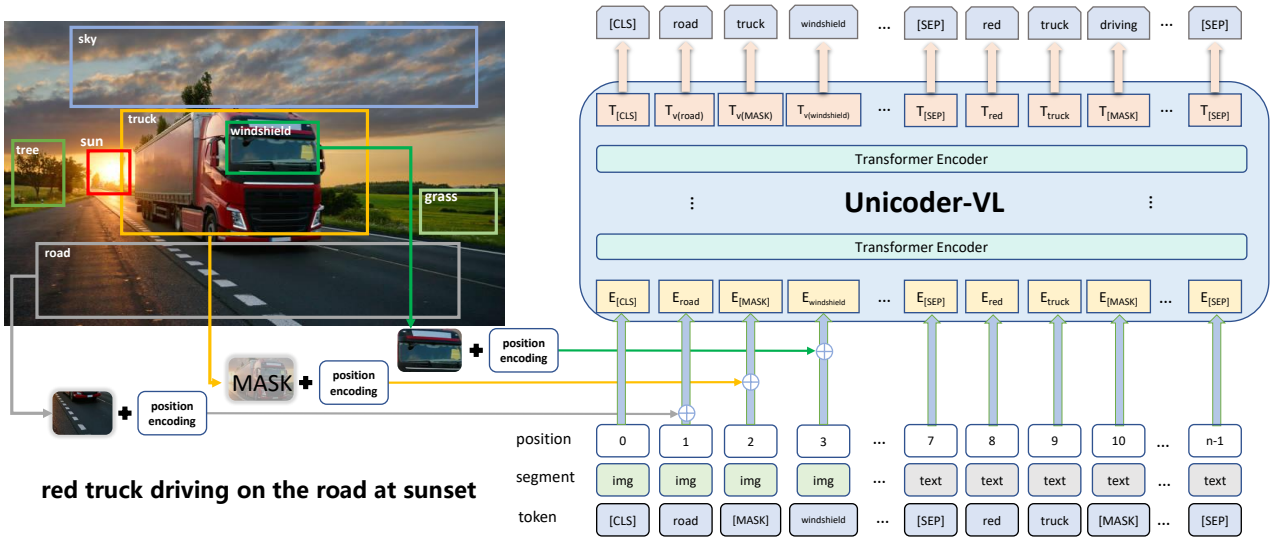

Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple transformer encoders which are used to learn viusal and linguistic representation jointly.

tionships between visual and linguistic contents. We fine-tune our pre-trained model to image-text retrieval 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 evaluate it on more cross-modal tasks soon, such as image captioning (Anderson et al. 2018), VQA (Antol et al. 2015), visual commonsense reasoning (Zellers et al. 2019), etc.

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) and XLNet (Yang et al., 2019), and 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.

ViLBERT (Lu et al. 2019), concurrently with our work, also describes a BERT-like architecture to jointly learn visual-linguistic representation. 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. However, we find using vanilla BERT structure, Unicoder-VL can outperform VilBERT significantly on both image retrieval and sentence retrieval tasks. We will verify this on more cross-modal tasks soon, including visual commonsense reasoning and visual QA.

Model

In this section, we first briefly summarize the original BERT model, and then 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.

Linguistic Representation. For each token in the input language sequence, its representation is a sum of token embedding, position embedding and segmentation embedding, which aims to differentiate linguistic contents and visual contents. We tokenize each input text into WordPieces (Wu et al. 2016) following the standard text pre-processing method of BERT. We also use the same vocabulary provided by BERT, which contains 30,522 tokens.

Image Representation. For each input image, we first detect objects from it using Detectron (Girshick et al. 2018), which is a pre-trained Faster R-CNN model. Here, top 100 objects with highest confidence scores are selected, each of which is represented as a vector computed by mean-pooling the last-layer convolutional feature of its region of interest. We also keep the predicted label of each detected object, which will be used in the object label prediction task.

Next, we represent the position each detected object as a 4-D vector, which is composed of the normalized bottom-left and top-right coordinates and contains both position and size information. Last, we concatenate the feature vector and position vector of each object, and transform it into another vector by linear projection, to make sure the dimensions of linguistic tokens and visual tokens are identical.

Multi-layer Transformer. The input vector sequence of Unicoder-VL consists of the vector sequence of linguistic tokens and the vector sequence of visual tokens. Similar to BERT, we add the special token `[CLS]` in the first position, and add another special token `[SEP]` between linguistic and visual tokens. After multiple self-attention layers, both linguistic tokens and visual tokens interact with each other very well and Unicoder-VL outputs the final representations.

Pre-training Tasks. We propose three tasks when doing the cross-modal pre-training: masked word prediction, masked object label prediction and visual-linguistic matching. Unlike VideoBERT (Sun et al. 2019), we do not use the image-only inputs (single modality).

Masked word prediction. For the linguistic part of the input, we mask the WordPiece tokens in each sequence at random, using the same procedure as BERT masked language model task. We omit the details and encourage to refer the original paper (Devlin et al. 2018).

Masked object label prediction. When processing the visual part, we replace the object feature vector with a random vector 80% of the time, we replace the object feature vector with a randomly chosen object feature from another image in 10% of the time and keep the object feature unchanged in the left 10% time. We simply take the label predicted by the same detection model as the object label and predict it using the final hidden state of the masked object feature.

Visual-linguistic matching. We take the final hidden state of the `[CLS]` token to predict whether the linguistic sentence is semantically matched with the visual content. Like original setting from BERT, we sample one random image as negative case and classify whether the image matches the linguistic sentence. Overall, we have three training regimes corresponding to the image-text inputs.

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.

Dataset

Pre-training Dataset. Deep learning models, in both visual and linguistic domains, have consistently demonstrated dramatic gains in performance with increasingly large datasets. Conceptual Captions dataset (Sharma et al. 2018) contains about 3.3M image and description pairs harvested from the web, which are very suitable for our cross-modal pre-training. Due to some broken urls, we finally use about 3M image-description pairs. Since the raw descriptions come from the Alt-text HTML attribute, some images and descriptions are not as relevant as some image caption dataset annotated by human like MSCOCO Caption and it is a noisy indicator of semantic relatedness.

Evaluation Dataset. The two datasets and their corresponding experimental protocols are introduced 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, in which each image is annotated with five caption sentences. 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.

Objective Function of Task-specific Fine-tuning.

Inputs of fine-tuning share the same data preprocessing procedures with pre-training. Note that we do not mask word and object in the fine-tuning stage. Similar to BERT sentence-pair classification tasks, we take the final hidden state corresponding to `[CLS]` token is used as the aggregate sequence representation for classification tasks. The only new parameters added during fine-tuning are from a classification layer. For image-text retrieval task, we take the output of the classification layer as matching score between image and text. We

propose two image-text matching tasks: image-to-text (i2t), text-to-image (t2i). We use triplet loss. When calculating triplet loss, we maximize the margin of positive and negative samples after generating the similarity score between two input modalities:

$$\mathcal{L}_{rank} = \sum_{y^- \in \mathcal{N}_y} \{\max[0, \gamma - s(x, y) + s(x, y^-)]\} \quad (1)$$

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

$$s(x, y) = W_o O_{[CLS]} + b_o \quad (2)$$

$O_{[CLS]}$ means the final state output of the `[CLS]` token. W_o is a linear projection layer.

In this study, we focus on the hardest negatives in every sampled examples, following (Faghri et al. 2017). For a positive pair (I, T), the hardest negatives are given by $I_h^- = \arg\max_{i \neq I} s(i, T)$ and $T_h^- = \arg\max_{t \neq I} s(I, t)$. Finally, we merge these ranking constraints into one loss function:

$$L_{hard} = \lambda_1 \sum_{I,T} \mathcal{L}_{hard}(I, T) + \lambda_2 \sum_{I,T} \mathcal{L}_{hard}(T, I) \quad (3)$$

where I is the image set and T is the captions of all images.

Implementation Details.

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 (including 100 objects and three special tokens). We use pre-trained parameters from BERT-bases, which is pre-trained on text data only.

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 in Section 3.3, using the ADAM optimizer with learning rate of 5e-5 with 20000 warmup steps, and a batch size of 192 with gradient accumulation (every 4 steps). We use float16 operations to speed up training and to reduce the memory usage of our models. The pre-training process takes more than 3 days.

During fine-tuning on image-text retrieval, we sample 3 negative cases in each matching tasks. We use $\gamma = 0.5$, $\lambda_1 = 1.0$, $\lambda_2 = 1.5$, $\lambda_3 = 0.05$ as the hyper-parameters of loss function. The optimizer is Adam and learning rate is set as 2e-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. The MSCOCO fine-tuning process takes 5 days. The Flickr30k fine-tuning takes more than 3 days.

Evaluation Results

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.

Results on MSCOCO Dataset. The experimental results on the MSCOCO dataset are shown in 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.

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 1K test set obtains 6.1% and 6.9% absolute improvements against the PFAN approach on sentence retrieval task and image retrieval task, respectively. For 5K test set, we can also significantly outperform all baselines on these two tasks.

Results on Flickr30k Dataset. The experimental results on the Filckr30K dataset is listed in Tab 2.

Both the zero-shot setting and the task-specific train setting show similar trends compared with results on MSCOCO. With pre-training and task-specific fine-tuning, Unicoder-VL achieves new state-of-the-art performance and yield a result of 82.3% and 68.3% on R@1 for sentence retrieval and image retrieval, respectively. Compared with PFAN, we achieve absolute boost of 12.3% on R@1 for sentence retrieval and 18.3% on R@1 for image retrieval.

We also compare Unicoder-VL with VilBERT (Lu et al., 2019) in the image retrieval setting. 10.1 points improvements show the superiority of Unicoder-VL.

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 viusal 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.

The results of Unicoder-VL outperform all the methods without jointly pre-training (acturally viusal 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 object featues 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.

In addition, The visual features using by VilBERT(Lu et al. 2019) are different from ours. So the results of retrieval task may be not comparable. However, we assume that for

Methods	Sentence Retrieval			Image Retrieval		
	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
m-CNN (Ma et al. 2015)	42.8	73.1	84.1	32.6	68.6	82.8
DSPE (Wang, Li, and Lazebnik 2016)	50.1	79.7	89.2	39.6	75.2	86.9
VSE++ (Faghri et al. 2017)	64.7	-	95.9	52.0	-	92.0
DPC (Zheng et al. 2017)	65.6	89.8	95.5	47.1	79.9	90.0
SCO (Huang et al. 2018)	69.9	92.9	97.5	56.7	87.5	94.8
SCAN (Lee et al. 2018)	72.7	94.8	98.4	58.8	88.4	94.8
SCG (Shi et al. 2019)	76.6	96.3	99.2	61.4	88.9	95.1
PFAN (Wang et al. 2019)	76.5	96.3	99.0	61.6	89.6	95.2
Unicoder-VL (zero-shot)	43.7	75.6	85.3	32.5	65.7	79.4
Unicoder-VL (task-specific train)	75.1	94.3	97.8	63.9	91.6	96.5
Unicoder-VL (pre-train + fine-tune)	**82.6**	**96.6**	**99.3**	**68.5**	**92.7**	**96.9**
5K Test set						
VSE++ (Faghri et al. 2017)	41.3	-	81.2	30.3	-	72.4
DPC (Zheng et al. 2017)	41.2	70.5	81.1	25.3	53.4	66.4
SCO (Huang et al. 2018)	42.8	72.3	83.0	33.1	62.9	75.5
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
Unicoder-VL (pre-train + fine-tune)	**59.6**	**85.1**	91.8	**44.5**	**74.4**	**84.0**

Table 1: Evaluation Results on MSCOCO testing set.

Methods	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (Karpathy and Fei-Fei 2015)	22.2	48.2	61.4	15.2	37.7	50.5
m-CNN (Ma et al. 2015)	33.6	64.1	74.9	26.2	56.3	69.6
DSPE (Wang, Li, and Lazebnik 2016)	40.3	68.9	79.9	29.7	60.1	72.1
VSE++ (Faghri et al. 2017)	52.9	79.1	87.2	39.6	69.6	79.5
DPC (Zheng et al. 2017)	55.6	81.9	89.5	39.1	69.2	80.9
SCO (Huang et al. 2018)	55.5	82.0	89.3	41.1	70.5	80.1
SCAN (Lee et al. 2018)	67.4	90.3	95.8	48.6	77.7	85.2
SCG (Shi et al. 2019)	71.8	90.8	94.8	49.3	76.4	85.6
PFAN (Wang et al. 2019)	70.0	91.8	95.0	50.4	78.7	86.1
VilBERT (Lu et al. 2019)	-	-	-	58.2	84.9	91.5
Unicoder-VL (zero-shot)	61.6	84.8	90.1	42.4	71.8	81.5
Unicoder-VL (task-specific train)	73.0	89.0	94.1	57.8	82.2	88.9
Unicoder-VL (pre-train + fine-tune)	**82.3**	**95.1**	**97.8**	**68.3**	**90.3**	**94.6**

Table 2: Evaluation Results on Flickr30k testing set.

Methods	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL (3-layer)	60.1	72.3	86.5	49.4	70.9	85.5
Unicoder-VL (6-layer)	72.4	93.1	96.3	58.1	83.4	90.2
Unicoder-VL (12-layer)	82.3	95.1	97.8	68.3	90.3	94.6

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. We found that the larger model boost the image-text retrieval tasks.

Methods	Visual	Sentence Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL	fix ResNeXt	4.0	14.1	22.9	4.3	15.5	26.1
Unicoder-VL	ft ResNeXt	4.7	18.5	28.5	5.6	18.8	30.9
Unicoder-VL	Faster R-CNN, 36 boxes	66.1	93.2	96.9	57.8	86.7	93.8
Unicoder-VL	Faster R-CNN, 100 boxes	82.6	96.6	99.3	68.5	92.7	96.9

Table 4: Comparisons of different image featuers, where "fixed" and "ft" refers to no fine-tuning and fine-tuning the image encoder, respectively.. All of these experiments are fine-tuning on **MSCOCO** with pre-trained Unicoder-VL.

different down-stream cross-modal tasks, both the model structure and features matters. Meanwhile, many labels of our 100 boxes are not recognized (we set them as "background" token). We ignore them in the masked object prediction task. There may be better pre-training tasks for the visual part of inputs, especially for these unrecognized objects.

Ablation Studies

In this section, we perform ablation experiments in order to better understand the pre-training tasks and the model.

Effect of Model Size. We compare the results transferring from Unicoder-VL models of varying transformer encoder layers. We test our model with 3-layer, 6-layer and 12-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. Due to limit resources, we did not choose to use BERT-large to initialize our Unicoder-VL, but 24-layer model initializing with BERT-large is probably more powerful than 12-layer Unicoder-VL.

Effect of Image Feature. We study the difference of pixel-level features and object-level featuers. For pixel-level features, we use ResNeXt-101 (32$*$4d) as the image encoder and elect the output of last pooling layer as the visual features. The size of visual feature map output by image encoder is 7$*$7$*$2048 and we consider it as visual features from 49 regions of an image. Note that we only do the masked word prediction and visual-linguistic matching task in this setting. For object features, we extract different boxes as visual inputs.

We can observe from Tab 4 that object-level features from object detection are clearly better than pixel-level features because the object region is related to a specific semantic object while the pixel-level features are not. If the model cannot learn some dependent relations, it will damage the pre-trained weights. And more parameters by involving the ResNext make the convergence more difficult. We will explore more pre-train tasks and fusion methods suitable for pixel-level features. Meanwhile, more detected regions may help improve the image-text retrieval task, though some of the region cannot be recognized correctly.

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. This pre-training method is general and not limited to image-text matching. We do not see any reason preventing it from finding broader applications in cross-modal tasks. We will further extend our pre-training method to VQA, and other cross-modal tasks in the future. 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.

References

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*.

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.

Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollár, P.; and He, K. 2018. Detectron. `https://github.com/facebookresearch/detectron`.

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, Y.; Wu, Q.; Song, C.; and Wang, L. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6163–6171.

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.

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.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: 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.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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*.

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*.

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.

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*.

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.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; and Shen, Y.-D. 2017. Dual-path convolutional image-text embedding with instance loss. *arXiv preprint arXiv:1711.05535*.