
ERNIE-ViL: KNOWLEDGE ENHANCED VISION-LANGUAGE REPRESENTATIONS THROUGH SCENE GRAPH

Fei Yu* Jiji Tang* Weichong Yin Yu Sun Hao Tian Hua Wu Haifeng Wang

Baidu Inc., China

{yufei07, tangjiji, yinweichong, sunyu02, tianhao, wu_hua, wanghaifeng}@baidu.com

August 3, 2020

ABSTRACT

We propose a knowledge-enhanced approach, ERNIE-ViL, to learn joint representations of vision and language. ERNIE-ViL tries to construct the detailed semantic connections (objects, attributes of objects and relationships between objects in visual scenes) across vision and language, which are essential to vision-language cross-modal tasks. **Incorporating knowledge from scene graphs, ERNIE-ViL constructs Scene Graph Prediction tasks**, i.e., Object Prediction, Attribute Prediction and Relationship Prediction in the pre-training phase. More specifically, these prediction tasks are implemented by predicting nodes of different types in the scene graph parsed from the sentence. Thus, ERNIE-ViL can model the joint representation characterizing the alignments of the detailed semantics across vision and language. **Pre-trained on two large image-text alignment datasets (Conceptual Captions and SBU)**, ERNIE-ViL learns better and more robust joint representations. It achieves state-of-the-art performance on 5 vision-language downstream tasks after fine-tuning ERNIE-ViL. Furthermore, it ranked the 1st place on the VCR leader-board with an absolute improvement of 3.7%.

1 Introduction

Motivated by pre-trained models like BERT [1] and ERNIE [2] which have significantly improved the performance on many NLP tasks, researchers ([3] [4] [5] [6] [7] [8] [9] [10]) have noticed the importance of pre-training for vision-language tasks, e.g., Visual Question Answering (VQA) [11] and Visual Commonsense Reasoning (VCR) [12].

Existing vision-language pre-training methods attempt to learn joint representations through visual grounding tasks on large image-text datasets, including Masked Language Modelling based on randomly-masked sub-words and Image-Text Matching at the whole image/text level. However, based on randomly-masking and predicting the sub-words, current models do not distinguish common words and words describing the detailed semantics [13], e.g., objects("man", "boat"), attributes of objects("boat is white"), relationships between objects("man standing on boat").

These methods neglect the importance of constructing detailed semantic alignments across vision and language, therefore the trained models can not well represent fine-grained semantics required by some real-world scenes. As shown in Figure 1, the detailed semantics are essential to distinguish the listed scenes which differ in objects, attributes and relationships. Hence better joint vision-language representations should characterize detailed semantic alignments across the modalities.

Inspired by the knowledge masking strategy of ERNIE 1.0 [2] which aimed at learning more structured knowledge by masking phrases and named entities rather than individual sub-words, we propose ERNIE-ViL, which incorporates knowledge from scene graphs [13], to construct better and more robust representations for vision-language joint modelling. Through constructing Scene Graph Prediction pre-training tasks, ERNIE-ViL puts more emphasis on

*indicates equal contribution.

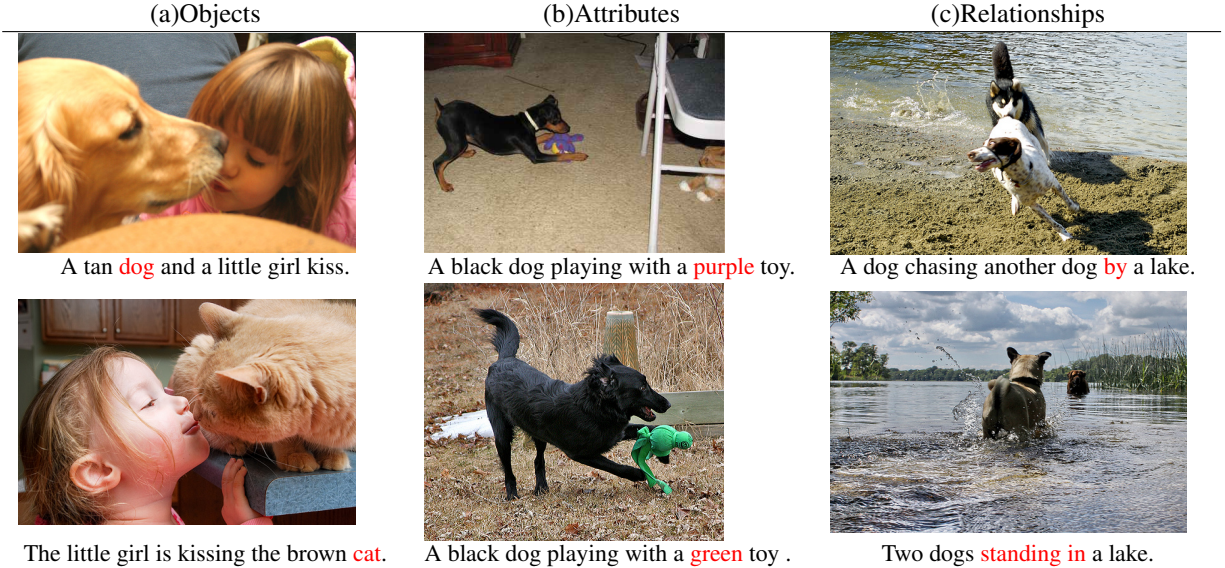


Figure 1: Scenes with subtle differences from Flickr30K dataset [17]. We show three pair of scenes, each of which differs respectively in (a)objects, (b)attributes and (c)relationships. The words with different information are marked in red.

detailed semantic alignments across vision and language. **Concretely, we implement these pre-training tasks based on predicting different types of nodes in the scene graph parsed from sentence.** The key insight lies in that during the pre-training phase, these tasks force the model to accurately predict the masked scene elements with the context of the observed information of both modalities, thus the model can learn the connections across the modalities. Through the Scene Graph Prediction tasks, ERNIE-ViL learns the detailed semantic alignments across vision-language modalities.

We pre-train ERNIE-ViL on two large commonly-used image-text out-of-domain datasets, Conceptual Captions [14] and SBU captions [15]. To evaluate the performance of ERNIE-ViL, we conduct experiments on various vision-language tasks, (1) visual question answering (VQA 2.0 [11]), (2) visual commonsense reasoning (VCR [12]) (3) region-to-phrase grounding (RefCOCO+ [16]) (4) image-text/text-image retrieval (Flickr30K [17]). On all these tasks, ERNIE-ViL obtains significant improvements compared to those methods pretrained on out-of-domain datasets. Specifically on the region-to-phrase grounding task, which needs detailed semantic alignments, we achieve an improvement of over 2.0% on both the testsets. And for fair comparison with the models pretrained on both out-of-domain and in-domain datasets, we further pretrain ERNIE-ViL on MS-COCO [18] and Visual-Genome [19] (in-domain datasets for downstream tasks). It achieves the state-of-the-art performances on all downstream tasks. Also we obtain the best single model performance and ranked the 1st place one the leaderboard with an absolute improvement of 3.7% on the Q->AR task compared to the former best model. Our code and pre-trained models will be publicly available.

Overall, our proposed method make three contributions:

- To the best of our knowledge, **ERNIE-ViL is the first work that introduces structure knowledge to enhance vision-language pre-training.**
- ERNIE-ViL constructs Scene Graph Prediction tasks during the pre-training of vision-language joint representations, putting more emphasis on the alignments of detailed semantics across modalities.
- ERNIE-ViL achieves state-of-the-art performances on 5 downstream cross-modal tasks and rank the 1st place on the VCR leaderboard.

2 Related Works

2.1 Cross-modal Pre-training

Inspired by text pre-training models [1], many cross-modal pre-training models for vision-language have been proposed. These researchers put their **efforts mainly on three aspects, which are model architecture, pre-training tasks and pre-training data.**

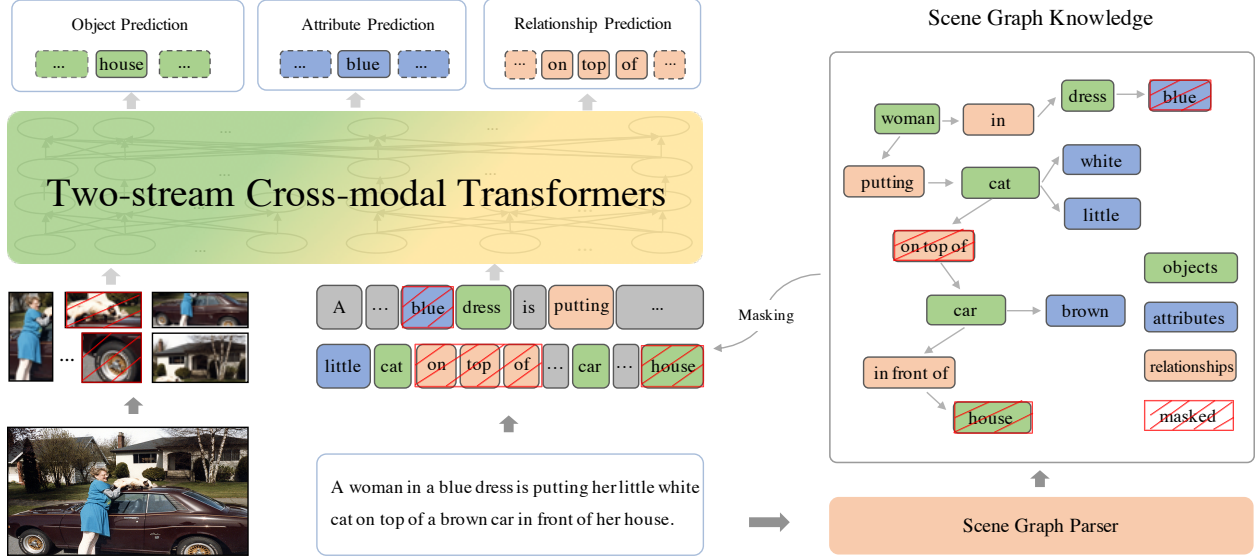


Figure 2: Illustration of Scene Graph Prediction for ERNIE-ViL. Given detected regions for image and token sequence for the text, ERNIE-ViL use a Two-stream Cross-modal Transformer network to model the joint vision-language representation. **Based on the scene graph parsed from the text using Scene Graph Parser**, we construct Object Prediction, Attribute Prediction and Relationship Prediction to learn the alignments of detailed semantics across modals.

Model Architecture These latest works are based on different variables of Transformers. Most of them ([5] [6] [7] [8] [9] [20]) use a uniform cross-modal Transformer modelling both image and text representations, while the others like ViLBERT [4] and LXMERT [21] are based on two-stream cross-modal Transformers, which bring more specific representations for image and text.

Pre-training Tasks Inspired by the pre-train tasks in text pre-training models, Masked Language Model and similar Masked Region Prediction tasks [4] are utilized in cross-modal pre-training. And similar to Next-Sentence Prediction, Image-Text Matching [4] [6] [10] task is also widely used. **However, only based on randomly masking and predicting sub-words, these methods do not distinguish the common words and words described the detailed semantics.** In this manner, the fine-grained semantic alignments across modalities cannot be well characterized in those learned joint representations.

Pre-training Data Unlike text pre-training models that can leverage tremendous natural language data, vision-language tasks require high-quality text-image aligned data that are hard to obtain. Conceptual Captions [14] and SBU Captions [15] are the most widely-used datasets for image-text pre-training, with 3.0M and 1.0M image descriptions respectively. These two datasets are **out-of-domain** for vision-language downstream tasks, while some existing works [10] [20] try to incorporate in-domain datasets, such as MS-COCO and Visual-Genome, which are highly correlation with downstream tasks.

2.2 Scene Graph

The scene graph contains structured knowledge of visual scenes, include the present objects, attributes of objects, and relationships between objects. As a beneficial prior knowledge describing the detailed semantics of the image and caption for the visual scene, scene graphs have led to many state-of-the-art models in image captioning [22], image retrieval [23], VQA [24] and image generation [25].

Various methods have been proposed to parse scene graphs from images [26] [27] and texts [28] [29] [30]. And scene graphs automatically parsed from the text have benefit several image-text multi-modal tasks. SPICE [29] proposed a new evaluation metric utilizing scene graphs parsed from captions for image captioning. UniVSE [23] improved the robustness in defending text-domain adversarial attacks for cross-domain tasks. SGAE [22] using scene graphs as internal structure bridging the gap of image and language modal to improving the performance of image captioning.

3 Approach

In this section, we will first introduce the model architecture of ERNIE-ViL. And then we will illustrate our newly-proposed **Scene Graph Prediction pretrain-training tasks**. Finally, pre-training with Scene Graph Prediction tasks in ERNIE-ViL will be introduced.

3.1 Model Architecture

The vision-language model aims at learning the joint representations that integrates information of both modalities and the alignments across the modalities. **The inputs to the joint model are usually a sentence and an image.**

3.1.1 Input Embedding

Given a sequence of words and an image, we first introduce the methods to embed the inputs to the feature space.

Sentence Embedding We adopt the similar word pre-processing method as BERT [1]. The input sentence is tokenized into sub-word tokens using WordPiece approach. Special tokens such as $[CLS]$ and $[SEP]$ are also added to the tokenized text sequence to make the text sequence $\{[CLS], w_1, \dots, w_T, [SEP]\}$. **The final embedding for each sub-word token is generated by combining its original word embedding, segment embedding and sequence position embedding.**

Image embedding For the image, we first use a pre-trained object detector to detect the salient image regions from the image. And the pooling features before multi-class classification layer are utilized as the region feature. We also encode the location features for each region via a 5-dimensional vector $(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2-y_1)(x_2-x_1)}{WH})$, where (x_1, y_1) and (x_2, y_2) denote the coordinate of the bottom-left and top-right corner while W and H are the width and height of the input image. We also add a special feature $[IMG]$ that denotes the representation the entire image (i.e. mean-pooled visual features with a spatial encoding corresponding to the entire image) to make the final region sequence $\{[IMG], v_1, \dots, v_t\}$.

3.1.2 Vision-Language Encoder

Given the embedding of image regions and the words for the sentence $\{[IMG], v_1, \dots, v_t, [CLS], w_1, \dots, w_T, [SEP]\}$, we use two-stream cross-modal Transformers to joint model the intra-modal and inter-modal representations. Following ViLBERT [4], ERNIE-ViL consists of two parallel BERT-style models operating over image regions and text segments.

The model outputs embeddings for each input of both the image and text: $h_{v_{i=1}^T}, h_{w_{i=1}^T}, h_{[IMG]}, h_{[CLS]}$ and $h_{[SEP]}$. We take $h_{[IMG]}$ and $h_{[CLS]}$ as the holistic image and text representations.

3.2 Scene Graph Prediction

As conditioned on both the sentence and the image, we could accurately reconstruct the objects(cat), attributes(white), and relationships (on top of) even if these elements are missing. However, only given the sentence, we could only reconstruct the elements with the same type as the origin tokens but without aligning them with image. **When the objects, attributes or relationships are masked in the sentence, the model cannot accurately reconstruct them without the help of the image.**

Scene graph encodes structured knowledge of visual scenes, including the present objects, attributes of objects, and relationships between objects, which are quite essential in differing scenes. Therefore, we construct Scene Graph Prediction tasks, i.e., Object Prediction, Attribute Prediction and Relationship Prediction Task. These tasks force the model to construct alignments across vision and language on more detailed semantics. Concretely, as shown in Figure 2, based on scene graph parsed from the text, we construct three prediction tasks according to the different node types in the scene graph, i.e. objects, attributes and relationships.

Scene graph parsing Given the text sentence \mathbf{w} , we parse it into a scene graph [29], which denotes as $G(\mathbf{w}) = \langle O(\mathbf{w}), E(\mathbf{w}), K(\mathbf{w}) \rangle$, where $O(\mathbf{w})$ is the set of object mentioned in \mathbf{w} , $E(\mathbf{w}) \subseteq O(\mathbf{w}) \times R(\mathbf{w}) \times O(\mathbf{w})$ is the set of hyper-edges representing relationships nodes between object nodes and $R(\mathbf{w})$ is the set of relationships mentioned in \mathbf{w} . $K(\mathbf{w}) \subseteq O(\mathbf{w}) \times A(\mathbf{w})$ is the set of attribute nodes associated with object nodes, where $A(\mathbf{w})$ is the set of attributes mentioned in \mathbf{w} . Scene graph describes the objects in more details with various attributes associated and relationships between objects. Thus integrating the knowledge of scene graph can benefit learning a more detailed joint representations for the vision-language. In this paper, the Scene Graph Parser provided by [29] is adopted to parse the

sentence: \mathbf{w}	A woman in a blue dress is putting her little white cat on top of a brown car in front of her house.
objects: $O(\mathbf{w})$	dress, woman, cat, car, house
relationships: $E(\mathbf{w})$	woman in dress, woman putting cat, cat on-top-of car, car in-front-of house
attributes: $K(\mathbf{w})$	blue dress, white cat, little cat, brown car

Table 1: A scene graph parsed from the caption of a visual scene.

text to scene graph. For a more intuitive understanding, we illustrate a specific case for the parsed scene graph from the text in Table 1.

3.2.1 Object Prediction

Objects are the dominant elements of the visual scenes, thus playing an important role in constructing the representation of semantics. Predicting the objects forces the model to build the vision-language connections at object level.

Firstly, for the all the objects nodes in the scene graph, we randomly select 30% of them to mask. And for each selected object node $O(\mathbf{w})$, we replace it with the special token $[MASK]$ in probability of 80%, another random token in probability of 10%, and keep it in probability of 10%. **Note that the objects are actually correspond to the sub-sequences of text in the sentence, therefore the object masking are implemented by masking the corresponding sub-sequences in the text.**

For Object Prediction, ERNIE-ViL tries to recover these masked object tokens, which is denoted as \mathbf{w}_{o_i} , based on the observation of their surrounding words \mathbf{w} and all image regions \mathbf{v} , by minimizing the negative log-likelihood:

$$\mathcal{L}_{obj}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \log(P(\mathbf{w}_{o_i} | \mathbf{w}_{\setminus o_i}, \mathbf{v})) \quad (1)$$

3.2.2 Attribute Prediction

Attributes characterize the detail information of the visual objects, such as color or shape of the objects, therefore encoding the detailed information in the visual scenes from another aspect.

Similarly, we randomly select 30% of all the attribute nodes in the scene graph, and the mask strategy is the same as that in Object Prediction. **Since the attribute nodes in the scene graph are attached to objects, we keep the associated object while masking out the attribute node $A(\mathbf{w})$ in each selected $K(\mathbf{w}) \subseteq O(\mathbf{w}) \times A(\mathbf{w})$.**

Given object words w_{o_i} in attribute pair $\langle w_{o_i}, w_{a_i} \rangle$, Attribute Prediction is to recover the masked tokens of attribute nodes, predicting the probability for each masked attribute word w_{a_i} . Based on the observation of the object tokens w_{o_i} , other surrounding words \mathbf{w} and all image regions \mathbf{v} , Attribute Prediction tries to minimize the negative log-likelihood:

$$\mathcal{L}_{attr}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \log(P(a_i | \mathbf{w}_{o_i}, \mathbf{w}_{\setminus a_i}, \mathbf{v})) \quad (2)$$

3.2.3 Relationship Prediction

Relationships describe the actions (semantic) or relative position (geometry) between the objects of the visual scenes, which contributes to distinguish scenes with same objects but different relationships.

When performing the mask strategy of selected relationship triplets $E(\mathbf{w}) \subseteq O(\mathbf{w}) \times R(\mathbf{w}) \times O(\mathbf{w})$, we keep the objects and mask out the relationship node $R(\mathbf{w})$. Thus, ERNIE-ViL constructs the Relationship Prediction task to learn the connections for the relationships across vision-language modalities. Specifically, given object tokens $\mathbf{w}_{o_{i1}}, \mathbf{w}_{o_{i2}}$ in relationship triplet $\langle \mathbf{w}_{o_{i1}}, \mathbf{w}_{r_i}, \mathbf{w}_{o_{i2}} \rangle$, this task tries to recover the masked relationship tokens, predicting the probability for each masked relation tokens \mathbf{w}_{r_i} . Thus the context for the prediction is the given object tokens $\mathbf{w}_{o_{i1}}, \mathbf{w}_{o_{i2}}$, other surrounding words from the text and all image regions \mathbf{v} :

$$\mathcal{L}_{rel}(\theta) = -E_{(\mathbf{w}, \mathbf{v}) \sim D} \log(P(\mathbf{w}_{r_i} | \mathbf{w}_{o_{i1}}, \mathbf{w}_{o_{i2}}, \mathbf{w}_{\setminus \mathbf{w}_{r_i}}, \mathbf{v})) \quad (3)$$

3.3 Pre-training with Scene Graph Prediction

Similiar to ViLBERT [4], ERNIE-ViL also adopts Masked Language Modelling (MLM) to capture the syntactic and lexical information in the text. Moreover, Masked Region Prediction and Image-text matching are utilized for visual modality and cross-modality respectively. The losses for these tasks are summed while pre-training.

	Text Stream				Image Stream				Cross Stream			
	L	H	A	F	L	H	A	F	L	H	A	F
Base	12	768	12	3072	6	1024	8	1024	6	1024	8	1024
Large	24	1024	16	4096	6	1024	16	4096	6	1024	16	4096

Table 2: Settings For ERNIE-ViL model. L : number of layers, H : hidden size, A : number of self-attention heads, F : feed-forward/filter size.

4 Experiments

4.1 Training ERNIE-ViL

Pre-training Data We use the Conceptual Captions (CC) dataset [14] and SBU dataset [15] as pre-training data. CC is a collection of 3.3 million image-caption pairs automatically scraped from alt-text enabled web images and SBU is a similar vision-language dataset which has 1.0 million image-caption pairs. Since some links had become broken by the time we downloaded the data, we only download about 3.0 million pairs for CC dataset and 0.8 million for SBU dataset. Notice that CC and SBU are image-caption pairs automatically collected from the web and have no intersections with the down-stream task datasets, thus act as out-of-domain datasets for training vision-language models.

Implementation Details For each image-text pair in the training, the pre-processing is performed as follows. For the image, we adopt Faster R-CNN [31] (with ResNet-101 [32] backbone) pre-trained on the Visual-Genome dataset to select salient image regions and extract region features. More specifically, regions with class detection probability exceeds a confidence threshold of 0.2 are selected and 10 to 36 boxes are kept. And for each kept region, the mean-pooled convolutional representation is used as the feature for it. For the text, we parse the scene graphs from the sentences using the Scene Graph Parser and adopt WordPieces to tokenize the sentence following BERT.

For the masking strategies, we randomly mask 15% tokens, 30 % scene graph nodes, and 15 % image regions. While for the token and region prediction tasks, only the item in the positive pairs will be predicted.

We train ERNIE-ViL on two model scale settings: ERNIE-ViL-base and ERNIE-ViL-large, which mainly differ in model depth for text stream. The detailed setting are shown in Table 2. We initialize the text stream with the parameters from ERNIE 2.0 model [33], and train ERNIE-VL with a total batch size of 512 for at least 500k steps on 8 V100 GPUs. And Adam optimizer with initial learning rates of $1e-4$ and a learning rate linear decay schedule is utilized.

4.2 Downstream Tasks

4.2.1 Visual Commonsense Reasoning (VCR)

The Visual Commonsense Reasoning (VCR) [12] task contains two sub-tasks: visual question answering ($Q \rightarrow A$) and answer justification ($QA \rightarrow R$), which are both multiple choice problems. The holistic setting ($Q \rightarrow AR$) requires both the chosen answer and the chosen rationale to be correct. The VCR dataset consists of 290k multiple choice QA problems derived from 110k movie scenes. In visual question answering ($Q \rightarrow A$) task, we concatenate the question and each candidate answer for the language modality and keep the image for the visual modality. **We take dot product of final hidden state of $h_{[CLS]}$ and $h_{[IMG]}$ to predict matching score for each answer semantically matched with the visual content with an additional FC layer. For the answer justification ($QA \rightarrow R$) task, we use the same setting as visual question answering ($Q \rightarrow A$) task.**

Similar with UNITER [10], a second-stage pre-training is utilized using VCR dataset. And then we fine-tune VCR model over 6 epochs with a batch size of 64 and initial learning rate of $2e-5$ which decays by 0.1 at the 2th and 4th epoch.

4.2.2 Visual Question Answering (VQA)

The VQA task requires answering natural language questions about images. VQA 2.0 dataset [11] contains 204k images and 1.1M questions about these images. Following [34], we treat VQA as a multi-label classification task – assigning a soft target score to each answer based on its relevancy to the 10 human answer responses. We take dot product of final hidden state of $h_{[CLS]}$ and $h_{[IMG]}$ to map this representation into 3,129 possible answers with an additional two layer MLP. The model is optimized with a binary cross-entropy loss on the soft target scores. Additional question-answer pairs from Visual Genome are used for data augmentation as in [10]. We fine-tune VQA model over 12 epochs with a batch size of 256 and initial learning rate of $1e-4$ which decays by 0.1 at the end of epoch 6 and epoch 9. At inference, we simply take a softmax.

4.2.3 Grounding Referring Expressions

The referring expression task is to localize an image region given a natural language reference. We evaluate the task on RefCOCO+ dataset [16]. In this paper, we use the bounding box proposals provided by [35] pre-trained on the MS-COCO dataset. We do the prediction for each region using its final hidden state h_{v_i} with an additional FC layer while each region i is labelled by computing the IoU with the ground truth box with a threshold of 0.5. We use a binary cross-entropy loss on the target label for each region and fine-tune RefCOCO+ model over 20 epochs with a batch size of 256, initial learning rate of $1e-4$ which decays by 0.2 at the end of epoch 5, 10, 15. At inference, we take the region with highest scoring as the prediction. The output of predicted bounding box is regarded as correct if the IoU between the predicted box and the ground truth box is higher than 0.5.

4.2.4 Image Retrieval & Text Retrieval

Caption-based image retrieval is the task of identifying an image from a pool given a caption describing its content. Flickr30K [17] contains 31,000 images collected from Flickr website where 5 captions are available for each image. Following the same split in [36], we use 1,000 images for validation and 1,000 images for testing and the rest for training.

We take dot product of final hidden state of $h_{[CLS]}$ and $h_{[IMG]}$ to predict matching score $s(w, v)$ for each text and image is matched with an additional FC layer. We utilize **circle loss** [37] with K random negative samples for each image-text pair. We set $K = 20$ for all settings. We trained 40 epochs on Flickr30K dataset with the initial learning rate $5e-6$ and decays at end of epoch 24 and epoch 32.

4.3 Results

We compare our pre-training ERNIE-ViL model against other cross-modal pre-training models. As shown in Table 3, with scene graph knowledge-enhanced, ERNIE-ViL achieves state-of-the-art results on all downstream tasks.

Pre-trained on the same out-of-domain datasets (CC, SBU), ERNIE-ViL acquires significant improvements on VCR, Image Retrieval and Text Retrieval compared to Unicoder-VL [5]. Specifically, its absolute improvement of 3.60% of R@1 for Image Retrieval and 2.50% of R@1 for Text Retrieval on Flickr30K demonstrates the effectiveness of detailed semantic alignments across vision and language. As compared to ViLBERT which uses the same two-stream cross-modal Transformers architecture, ERNIE-ViL obtains better results on all downstream tasks. Notice that ERNIE-ViL is pre-trained only on out-of-domain datasets, therefore for the fair comparison with those models pretrained with out-of-domain and in-domain datasets, we further pre-train ERNIE-ViL with in-domain datasets (Visual-Genome, MS-COCO). As illustrated in Table 3, ERNIE-ViL-large achieves better downstream task performances on 5 tasks compared to UNITER, 12-in-1 [38], OSCAR [39] and VILLA [40].

4.4 Analysis

To validate the effectiveness of incorporating knowledge from scene graph, we conduct the language cloze test conditioned on the visual modality.

In the cloze test, language tokens represent detailed semantics (objects, attributes and relationships) are masked from the text and the model is required to infer them with the context from both text and image. To build the dataset, we sampled 15,000 image-text pairs from Flickr30K dataset and in total each of 5,000 object tokens, attributes and relationships are selected. And for the prediction, acc@1 and acc@5 are adopted as the evaluation metric. The comparison of prediction results between baseline model, which is pre-trained without Masked Scene Graph prediction task, and proposed ERNIE-ViL is illustrated in Table 4. An absolute improvement acc@1 of 2.12% for objects, 1.31% for relationships and 6.00% for attributes demonstrates that ERNIE-ViL learned better alignments for detailed semantics across modalities.

Moreover, we also illustrate some cases in Table 5, and the top 5 possible predictions are shown in the right columns. As in case 1-5, the baseline model cannot make the right predictions as it didn't learn accurate alignments of detailed semantics without distinguishing common words and detailed semantics while pre-training. While in case 6, the baseline model can predict the reasonable tokens but with lower confidence. However, ERNIE-ViL may also predict incorrect tokens in case 7-8 due to the fact that the detailed semantics ("yellow", "brown" in case 7 and "dog", "animal" in case 8) in the visual space are quite similar.

Models		VCR			RefCOCO+		
		Q→A	QA→R	Q→AR	val	testA	testB
Out-of-domain	ViLBERT-base	72.42 (73.3)	74.47 (74.6)	54.04 (54.8)	72.34	78.52	62.61
	Unicoder-VL-base	72.6 (73.4)	74.5 (74.4)	54.4 (54.9)	-	-	-
	VLBERT-base	73.8 (-)	74.4 (-)	55.2 (-)	71.60	77.72	60.99
	UNITER-base	-	-	-	72.78	-	-
	VLBERT-large	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	72.59	78.57	62.30
	ERNIE-ViL-base	74.37 (77.0)	79.65 (80.3)	61.24 (62.1)	74.02	80.33	64.74
	ERNIE-ViL-large	78.52(79.2)	83.37(83.5)	65.81(66.3)	74.24	80.97	64.70
Out-of-domain + in-domain	UNITER-base	74.56 (75.0)	77.03 (77.2)	57.76 (58.2)	75.31	81.30	65.58
	VILLA-base	75.54 (76.4)	78.78 (79.1)	59.75 (60.6)	76.05	81.65	65.70
	UNITER-large	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	75.90	81.45	66.70
	VILLA-large	78.45 (78.9)	82.57 (82.8)	65.18 (65.7)	76.17	81.54	66.84
	ERNIE-ViL-large	78.62 (-)	83.42 (-)	65.95(-)	75.95	82.07	66.88

Models		VQA		IR-Flickr30K			TR-Flickr30K		
		test-dev	test-std	R@1	R@5	R@10	R@1	R@5	R@10
Out-of-domain	ViLBERT-base	70.55	70.92	58.20	84.90	91.52	-	-	-
	Unicoder-VL-base	-	-	71.50	90.90	94.90	86.20	96.30	99.00
	VLBERT-base	71.16	-	-	-	-	-	-	-
	UNITER-base	71.56	-	-	-	-	-	-	-
	VLBERT-large	71.79	72.22	-	-	-	-	-	-
	ERNIE-ViL-base	72.62	72.85	74.44	92.72	95.94	86.70	97.80	99.00
	ERNIE-ViL-large	73.78	73.96	75.10	93.42	96.26	88.70	97.30	99.10
Out-of-domain + in-domain	UNITER-base	72.70	72.91	72.52	92.36	96.08	85.90	97.10	98.80
	OSCAR-base	73.16	73.61	-	-	-	-	-	-
	VILLA-base	73.59	73.67	74.74	92.86	95.82	86.60	97.90	99.20
	12-in-1-base	73.15	-	67.90	-	-	-	-	-
	UNITER-large	73.82	74.02	75.56	94.08	96.76	87.30	98.00	99.20
	OSCAR-large	73.44	73.82	-	-	-	-	-	-
	VILLA-large	74.69	74.87	76.26	94.24	96.84	87.90	97.50	98.80
	ERNIE-ViL-large	74.75	74.93	76.70	93.58	96.44	88.10	98.00	99.20

Table 3: Results on downstream V+L tasks for ERNIE-ViL model, compared with previous state-of-the-art pre-trained models. IR: Image Retrieval. TR: Text Retrieval.

	objects		attributes		relationships		overall	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Baseline model	54.74	78.30	40.08	61.94	48.01	67.19	47.59	69.21
ERNIE-ViL	56.86	79.54	46.08	68.76	49.32	69.39	50.80	72.67

Table 4: Cloze Test for ERNIE-ViL. Baseline model denotes pre-training without Scene Graph Prediction Tasks.

5 Conclusion

We proposed ERNIE-ViL approach to learn the joint representations of vision and language. In addition to conventional MLM for cross-modal pre-training, we introduce Scene graph Prediction to characterize the detailed semantic alignments across vision and language. Experiment results on various downstream tasks demonstrate the improvements of incorporating knowledge from scene graph during cross-modal pre-training. For future work, scene graph extracted from images could also be incorporated into cross-modal pre-training. Moreover, Graph Neural Networks to integrate more structured knowledge could be considered as well.

References

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [2] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*,


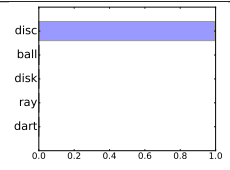
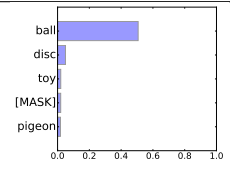

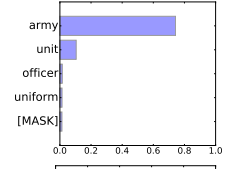
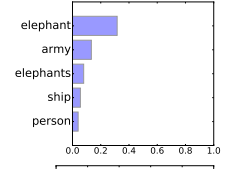

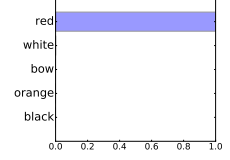
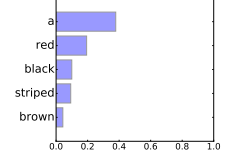

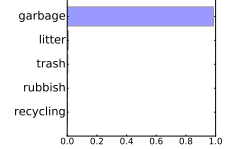
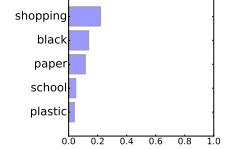

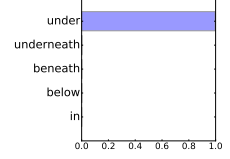
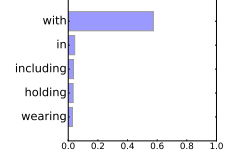

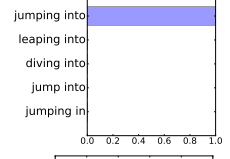
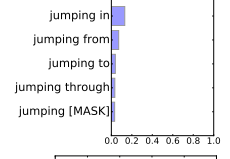

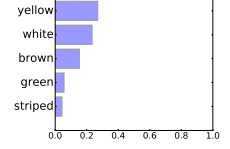
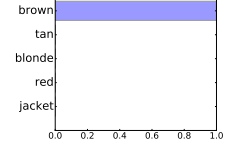
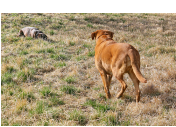
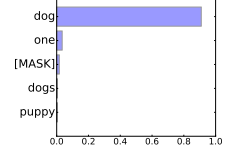
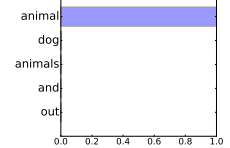
	Image	Text	ERNIE-ViL	Baseline
1		a black dog about to catch a flying disc .		
2		here is a picture of an army standing in one row and on an island .		
3		a man with a white shirt and red tie is talking to another man in a kitchen .		
4		two teen boys in school clothes are walking with something in a garbage bag .		
5		five children are on a carnival ride under a clown face .		
6		two dolphins jumping into the water .		
7		a man in a brown shirt is cutting a piece of cake .		
8		a brown dog walks towards another animal hiding in the grass .		

Table 5: Effectiveness of Scene Graph Predictions Strategies for ERINE-ViL. Each sample represents an image-text pair with the masked token colored in red.

2019.

- [3] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019.
- [4] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

- [5] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019.
- [6] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- [7] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.
- [8] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.
- [9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.
- [11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [12] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.
- [13] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.
- [14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [15] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011.
- [16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.
- [17] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.
- [21] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [22] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.
- [23] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019.
- [24] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*, 2019.
- [25] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

- [26] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.
- [27] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017.
- [28] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.
- [29] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.
- [30] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. *arXiv preprint arXiv:1803.09189*, 2018.
- [31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [33] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*, 2019.
- [34] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086, 2018.
- [35] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.
- [36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.
- [37] Yifan Sun, Changmao Cheng, Yuhang Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. *arXiv preprint arXiv:2002.10857*, 2020.
- [38] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [39] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020.
- [40] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.