ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration

Yuhao Cui¹, Zhou Yu*, Chunqi Wang², Zhongzhou Zhao², Ji Zhang², Meng Wang³, Jun Yu¹
¹School of Computer Science and Technology, Hangzhou Dianzi University, China
²Alibaba Group, China
³School of Computer Science and Information Engineering, Hefei University of Technology, China

{cuiyh,yuz,yujun}@hdu.edu.cn,{shiyan.wcq,zhongzhou.zhao,zj122146}@alibaba-inc.com,eric.mengwang@gmail.com

ABSTRACT

Vision-and-language pretraining (VLP) aims to learn generic multimodal representations from image-text pairs. While various successful attempts have been proposed, learning fine-grained semantic alignments between image-text pairs plays a key role in their approaches. Nevertheless, most existing VLP approaches have not fully utilized the intrinsic knowledge within the image-text pairs, which limits the effectiveness of the learned alignments and further restricts the performance of their models. To this end, we introduce a new VLP method called ROSITA, which integrates the cROSS-S- and InTRA-modal knowledge in a unified scene graph to enhance the semantic alignments. Specifically, we introduce a novel structural knowledge masking (SKM) strategy to use the scene graph structure as a priori to perform masked language (region) modeling, which enhances the semantic alignments by eliminating the interference information within and across modalities. Extensive ablation studies and comprehensive analysis verify the effectiveness of ROSITA in semantic alignments. Pretrained with both in-domain and out-of-domain datasets, ROSITA significantly outperforms existing state-of-the-art VLP methods on three typical vision-and-language tasks over six benchmark datasets.

CCS CONCEPTS
• Computing methodologies → Multi-task learning.

KEYWORDS
vision-and-language; deep learning; knowledge-enhanced learning; multimodal pretraining

1 INTRODUCTION

Motivated by the success of the pretrain-then-finetune paradigm of BERT in natural language understanding [8], there has been an increasing interest in developing vision-and-language pretraining (VLP) models [7, 23, 27, 38] to address a wide range of vision-and-language (V+L) tasks. In particular, these approaches first pretrain transformer-based models on large image-text corpus to learn task-agnostic representations, and then finetune the models on downstream V+L tasks, e.g., visual question answering [50, 51], image text retrieval [18, 31], and referring expression comprehension [16, 52]. Compared to earlier methods that are only adapted to one V+L task [46, 47, 49], VLP models is generalizable across multiple tasks and also achieves significantly better performance on respective tasks.

Learning fine-grained semantic alignments between image regions and text words plays a key role in V+L tasks. However, manually annotating such dense alignment between regions and words is expensive and is unrealistic under the large-scale scenario. Therefore, most existing VLP approaches [7, 21, 27] use a weakly-supervised learning strategy to model the alignments implicitly. Taking the image regions and text words as inputs, they adopt multi-layer Transformers [39] as their backbones to learn fine-grained semantic alignments.
We briefly review previous studies on unimodal pretraining and propose a novel approach to selectively mask cross- and intra-modal contexts while enhancing the semantic alignments by enforcing the model to acquire accurate intrusion information within and across modalities. This strategy effectively eliminates the interference between the anchor object and its intra-modal contexts (e.g., objects along with their attributes and relations) in the scene graph in its pretraining objectives. In terms of knowledge source, both of them use the intra-modal knowledge from a single modality to enhance the semantic alignments: OSCAR models the intra-modal knowledge in the image modality while ERNIE-ViL models the intra-modal knowledge in the text modality. The success of the two methods above raises a question: Is it possible to utilize the intra-modal knowledge from both modalities along with the cross-modal knowledge to further enhance the semantic alignments?

In this paper, we present a new VLP method called ROSITA, which encodes the cross-modal and intra-modal knowledge simultaneously in a unified scene graph. As shown in Figure 1, the graph consists of a set of knowledge entries, where each entry corresponds to a scene graph constructed from text and image modalities. The intra-modal knowledge refers to the relationships between the anchor object and its intra-modal contexts (e.g., spatially related regions or contextually related words). The cross-modal knowledge corresponds to the relationships between the anchor object and its semantically similar objects from the opposite modality (e.g., the region predicted as "grass" and the word "steppe").

Although we have obtained a set of knowledge entries, how to effectively use them to enhance semantic alignments is still nontrivial. We propose a novel structural knowledge masking (SKM) strategy that can be seamlessly integrated with the masked language (region) modeling tasks, which are commonly used in existing VLP methods [7, 27]. The SKM strategy first masks the anchor object while selectively masking its cross- and intra-modal contexts in a knowledge entry. This strategy effectively eliminates the interference information within and across modalities and enhances the semantic alignments by enforcing the model to acquire accurate information from the opposite modality.

The contributions of this work are three-fold:

1. We present a new VLP method ROSITA, which incorporates cross- and intra-modal knowledge simultaneously to enhance the semantic alignments across different modalities.
2. We introduce a novel structural knowledge masking strategy to use the scene graph structure as a prior to be integrated with the commonly used masked language (region) modeling tasks in existing VLP methods.
3. We achieve the best results on three typical V+L tasks over six benchmark datasets, outperforming existing state-of-the-art VLP methods.

# 2 RELATED WORK

We briefly review previous studies on unimodal pretraining and vision-and-language pretraining, especially those studies on knowledge enhanced pretraining.

## Unimodal Pretraining

The pretraining technique has been widely used in computer vision (CV) tasks. Deep convolutional neural networks like VGGNet [35] or ResNet [12] pretrained on ImageNet can well generalize to various downstream tasks [11, 26, 33]. In contrast to CV tasks, the popularization of pretraining in the natural language processing (NLP) community is relatively late. Based on the multi-layer Transformer architecture [39], many famous pretraining approaches (e.g., BERT [8], GPT [32], and XLNet [42]) have been put forward. Different from the supervised pretraining paradigm in CV tasks, the pretraining paradigm in NLP tasks is self-supervised that aims to train a model to predict words based on their contexts without introducing human annotations. In particular, BERT introduces a novel masking language modeling (MLM) task that randomly masks the input words and predicts these masked words based on their contexts. This MLM strategy is naturally inherited by the VLP methods.

## Vision-and-Language Pretraining (VLP)

Different from the purely self-supervised paradigm in NLP tasks, VLP models are pretrained on large-scale paired image-text corpus, e.g., image captioning datasets like [6, 29, 34]. Mirroring the success of BERT, recent studies naturally extend its framework to the vision-and-language domain to pretrain VLP models for a wide range of V+L tasks [7, 13, 23, 27, 38, 45, 53]. ViLBERT [27] and LXMERT [38] are two pioneering works in this field, where the two-stream architectures are adopted to encode the image features and textual features with two separate Transformers and then perform multimodal fusion via a third Transformer. Recent works tend to use the single-stream architectures, where the multimodal features are directly fused using one Transformer [7, 19, 21, 36]. Moreover, other techniques like knowledge integration [23, 45], multilingual enhancement [55], contrastive learning [22], and adversarial training [9] are introduced to further improve the performance of the pretrained models.

## Knowledge-Enhanced Pretraining

Incorporating prior knowledge (e.g., external knowledge graph) to enhance model pretraining has been investigated earlier by two ERNIE methods [37, 54] and widely explored in recent years [25, 40, 41]. The introduced prior knowledge enables the model to better understand the syntactic and semantic structure of the text, thus facilitating model pretraining by an improved structural MLM task. In the VLP task, prior knowledge can be acquired from both the image and text modalities. ERNIE-ViL constructs a scene graph from text and puts more emphasis on the discovered keywords [45]. OSCAR exploits the predicted tags of image regions to enhance the semantic alignment across the two modalities [23]. A concurrent work UC2 utilizes off-the-shelf machine translation model to construct aligned multilingual dataset for texts and regard this extra information as prior knowledge to enhance the learning of cross-modal semantic alignment [55].

Despite the success of these knowledge-enhanced VLP methods, they only utilize the intra-modal knowledge from a single modality, which restricts their effectiveness in learning semantic alignments.

To the best of our knowledge, our ROSITA is the first VLP method to integrate the cross-modal and intra-modal knowledge simultaneously in order to enhance the learning of semantic alignments across different modalities.


3 KNOWLEDGE EXTRACTION

In this section, we introduce the procedure of extracting knowledge entries from an image-text pair. We first construct a unified graph to model the intra- and cross-modal knowledge from an image-text pair. On top of the established graph, we select anchor objects to obtain a set of knowledge entries. The process of knowledge extraction is illustrated in Figure 2.

Unified Scene Graph Construction. Given an image-text pair, we resort to a unified scene graph structure \( G = (V, E, S) \) to encode its intra- and cross-modal knowledge simultaneously [48]. The vertex set \( V \) includes the words and regions from the text and image, respectively. The edge set \( E \) and similarity set \( S \) contain pairwise relationships and their corresponding similarities between vertices (i.e., edge weights), respectively.

The intra-modal knowledge within the image and text are first represented as an image scene graph and a text scene graph, respectively. For the image scene graph, regions extracted from a pretrained object detector are considered as the vertices in \( V \). Inspired by [14, 20, 43], we calculate the similarity between each paired regions by their Intersection over Union (IoU) score. The region pairs with IoU scores larger than zero are considered to have edges in \( E \) and their IoU scores are regarded as their similarities in \( S \). For the text graph, we use an off-the-shelf scene graph parser provided by [1] to obtain a text scene graph from a text. The text scene graph explicitly encodes the keywords of objects, attributes, and relations found in the text while discarding the rest of uninformative words. These mentioned keywords in the scene graph are regarded as the vertices in \( V \). The word-word relationships in the scene graph (i.e., object-attribute or object-relation) correspond to the edges in \( E \). The similarity between two vertices is the co-occurrence frequencies of the referred object-attribute (or object-relation) pair calculated on the whole dataset. Since the similarity distributions of the image and text modalities may vary widely, we normalize the similarities within each modality, respectively.

As we have modeled the intra-modal knowledge in the graph, we further integrate cross-modal knowledge to align the image regions to their semantically related words. Since such cross-modal alignment supervision is not available, we establish pseudo semantic alignments between region-word pairs as follows. For the image regions, the predicted region tags are aligned to the object words with respect to their semantic similarities on words. We adopt a pretrained word embedding model [30] to calculate the pairwise similarities between object tags and object words. We set a minimum confidence threshold of 0.5 to the similarity scores to make a trade-off between precision and recall. The region-word pairs surpass the threshold will form cross-modal edges in \( E \) and their corresponding scores represent the similarities in \( S \).

Knowledge Representation. Based on the constructed unified scene graph \( G \), we illustrate the procedure of extracting knowledge entries from the scene graph in detail. Note that each knowledge entry is associated with an anchor object, we first select all possible anchor objects from the graph. We define an anchor object as the vertex (an image region or a text word) in the graph that is referred to by at least one cross-modal edge. Since the attribute and relation words are not directly connected to any image region, they cannot be anchor objects according to our definition.

After obtaining the anchor objects, we integrate the intra-modal knowledge and cross-modal knowledge in \( G \) to obtain a knowledge entry. Given an anchor object \( o \in V \), its corresponding knowledge entry is represented as a subgraph \( g(o) \subseteq G \) and is obtained by the

\[1\] We have tried to establish more fine-grained alignments to include the attribute words. However, the predicted attributes from image regions are too diverse that often fail to match the attribute words in the text.
union of three subgraphs of $G$ as follows:

$$ g(\mathbf{v}) = G_{\text{cross}}(\mathbf{v}) \cup G_{\text{intra}}(\mathbf{v}) \cup G_{\text{intra}}(G_{\text{cross}}(\mathbf{v})) $$  \hspace{1cm} (1)$$

where $G_{\text{cross}}(\mathbf{v})$ contains the relationships between $\mathbf{v}$ and its directly connected contexts by cross-modal edges. $G_{\text{intra}}(\mathbf{v})$ models the relationships between $\mathbf{v}$ and its directly connected contexts by intra-modal edges. $G_{\text{intra}}(G_{\text{cross}}(\mathbf{v}))$ includes the relationships between the vertices in $G_{\text{cross}}(\mathbf{v})$ and their corresponding intra-modal contexts. It is worth noting that the anchor object $\mathbf{v}$ can reach every vertex in $g(\mathbf{v})$ within two hops.

## 4 THE ROSITA FRAMEWORK

Based on the extracted knowledge entries from image-text pairs, we introduce the ROSITA framework in this section. We first describe the image and text feature representations and the network architecture. Then, we introduce a structural knowledge masking (SKM) strategy, which takes the knowledge entries as a priori to perform the masked language (region) modeling. Finally, we describe the whole pretraining objective with multi-task learning. The overall framework is illustrated in Figure 3.

### Image and Text Feature Representations

Following the commonly used strategy in existing VLP methods [6, 23], the input image is represented as a set of regional features, which are extracted from a Faster R-CNN model pretrained on Visual Genome [2]. More specifically, we extract $m$ regions with the highest confidence probabilities from the image. For the $i$-th region, it is represented as a visual feature $\mathbf{f}_i \in \mathbb{R}^{2048}$ and a positional feature $\mathbf{p}_i \in \mathbb{R}^{5}$ [45]. The two features are fused into a $d$-dimensional image representation $\mathbf{x}_i \in \mathbb{R}^d$ using two linear projections as follows:

$$ \mathbf{x}_i = W_f^T \mathbf{f}_i + W_p^T \mathbf{p}_i $$  \hspace{1cm} (2)

where $W_f \in \mathbb{R}^{2048 \times d}$ and $W_p \in \mathbb{R}^{5 \times d}$. Finally, the image is represented as a feature matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.

For its paired text, we adopt the word processing method similar to [8]. The input text is first tokenized into words and trimmed (or padded) to a maximum of $n$ words. Each word $\mathbf{w}_i$ and its index $i$ (i.e., the absolute position of $\mathbf{w}_i$ in the text) are projected to vectors by two individual embedding layers, then added to obtain the position-aware text representation $\mathbf{y}_i$ as follows:

$$ \mathbf{y}_i = \text{WordEmbed}(\mathbf{w}_i) + \text{IdxEmbed}(i) $$  \hspace{1cm} (3)

where $\mathbf{y}_i$ is $d$-dimensional to match the image representation. The text is finally represented as a feature matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$.

### Network Architecture

The image features $\mathbf{X} = [x_1, ..., x_m]$ and text features $\mathbf{Y} = [y_1, ..., y_n]$ are first concatenated before feeding to the network. We insert two special tokens to the concatenated features to obtain the multimodal input features $\mathbf{Z}$:

$$ \mathbf{Z} = \{ x_1, x_2, ..., x_m, [\text{SEP}], y_1, y_2, ..., y_n, [\text{CLS}] \} $$  \hspace{1cm} (4)

where the [SEP] token marks the boundary between the image and text features. The [CLS] token is used to predict whether the given image and text are paired or not.

The multimodal features $\mathbf{Z}$ are fed into a single-stream Transformer with $L$ layers [8]. Each layer consists of a multi-head self-attention (MSA) block and a feed-forward networks (FFN) block. The random masking based MLM and MRM tasks are not aware of the keywords and key regions to be aligned, their efficacy in the alignment learning is weak. To this end, we present an alternative structural knowledge masking (SKM) strategy to selectively mask the tokens referred to by the extracted knowledge entry. Accordingly, the MLM and MRM tasks are respectively modified to the SKMLM and SKMRM tasks to adapt to the SKM strategy.

Let an image be represented as a set of regions $\mathcal{R} = \{r_1, ..., r_m\}$ and a text be represented as a sequence of words $\mathcal{W} = \{w_1, ..., w_m\}$, we construct a unified scene graph $G$ on top of $\mathcal{R}$ and $\mathcal{W}$, and extract a set of knowledge entries from $G$. Let $g(\mathbf{v}) = g_i < \mathcal{V}, \mathcal{E}, \hat{\mathcal{S}} >$ be one of the knowledge entries, where $\mathbf{v}_i \in \mathcal{V}$ is the anchor object. The vertices are represented as $\hat{\mathbf{V}} = \{v_1, ..., v_{n_x}\}$ and the similarities between the vertices are represented as $\hat{\mathcal{S}} \in \mathbb{R}^{N \times N}$, where $N$ is the number of vertices in this entry.

The strategy of SKM is to deterministically mask the anchor object $\mathbf{v}_i$ while probabilistically masking its intra-modal contexts and cross-modal contexts with respect to the graph structure of the knowledge entry. Since the similarities between $\mathbf{v}_i$ and its contexts are different, we assign independent masking probabilities to each of the contexts with respect to their similarities to $\mathbf{v}_i$, rather than simply using an identical masking probability for all the contexts. To obtain the masking probabilities for the contexts, we introduce a masking strategy that satisfies the following principles: for the intra-modal contexts, a larger similarity score refers to a higher masking probability. For the cross-modal contexts, a larger score leads to a lower masking probability. The reasons behind this masking strategy will be explained hereinafter.

Note that not all contexts have direct connections to the anchor object. Therefore, we calculate the transmission probabilities $\mathbf{T} = [t_1, ..., t_N] \in [0, 1]^N$ from the anchor object $\mathbf{v}_i$ to its contexts in $\hat{\mathbf{V}}$.
based on the normalized similarities defined in $\hat{S}$. Since $v_i$ can reach all the vertices in $V$ within two hops, $T$ is defined as follows:

$$T = \frac{1}{2} \hat{S} \pi(i) + \frac{1}{2} \hat{S} \pi(i)$$ (6)

where $\pi(i) \in \{0, 1\}^N$ is a one-hot vector with the $i$-th element to be 1. The two terms correspond to the transmission probabilities between the anchor object $v_i$ and its one-hop and two-hop contexts, respectively. After that, we convert the transmission probabilities $T$ to the masking probabilities $P = \{p_{t_1}, ..., p_{t_N}\} \in \{0, 1\}^N$ using the following rules to satisfy our masking strategy above:

$$p_j = \begin{cases} 1, & \text{if } v_j \text{ is the anchor object} \\ \alpha t_j, & \text{if } v_j \text{ is within the intra-modal contexts} \\ (1 - \alpha)(1 - t_j), & \text{if } v_j \text{ is within the cross-modal contexts} \end{cases}$$ (7)

where $\alpha$ is a hyper-parameter to balance the transmission probabilities of the intra-modal and cross-modal contexts. $t_j$ and $p_j$ denote the transmission and masking probability of the vertex $v_j$, respectively.

Given a knowledge entry, we use the calculated masking probabilities to obtain two groups of mask indices $M_w$ and $M_r$, indicating the words and regions to be masked, respectively. The partially masked input features are passed through the network and then fed into the SKMLM and SKMRM tasks.

In particular, if the anchor point $v_i$ refers to an object word in the text, we resort to the SKMLM task to reconstruct the masked words $W_{M_w}$ as follows:

$$L_{SKMLM}(\theta) = -\mathbb{E}_{(W, R) \sim D} \log P_{\theta}(W_{M_w} | W_{\bar{M}_w}, R_{\bar{M}_r})$$ (8)

where $\theta$ is the trainable parameters. Each pair $(W, R)$ is sampled from the whole training set $D$. $W_{\bar{M}_w}$ and $R_{\bar{M}_r}$ refer to the remaining words in $W$ and the remaining regions in $R$ with excluding the masked tokens from their modalities, respectively.

Analogously, if $v_i$ refers to a region in the image, we resort to the SKMRM task to reconstruct the masked regions $R_{M_r}$ as follows:

$$L_{SKMRM}(\theta) = -\mathbb{E}_{(W, R) \sim D} f_{\theta}(R_{M_r} | W_{M_w}, R_{\bar{M}_r})$$ (9)

where $f_{\theta}(\cdot)$ refers to some loss functions. Similar to [7], we use the regression-based loss and classification-based loss joint.

The motivations of SKM can be explained as follows: (i) The intra-modal contexts may contain interference information (e.g., the word “sky” is frequently associated with an attribute “blue” and a visual object of “wheel” is usually within an object of “car”). Such interference information may leak out the semantics of the anchor object and reduce the difficulty of anchor object reconstruction, leading to a degradation of the pretrained model. Therefore, when masking an anchor object, its intra-modal contexts with high similarities will have high probabilities to be masked simultaneously. This operation reduces the risk of information leakage and enforces the model to acquire precise information from the opposite modality, which implicitly enhance the semantic alignments; (ii) The cross-modal contexts with low similarities may contain irrelevant or noisy information. Therefore, when masking an anchor object, its cross-modal contexts with high similarities will have low probabilities to be masked at the same time, which explicitly excludes potential noise thus benefiting the semantic alignments. As a result, the synergy of the masking operations above significantly facilitates the semantic alignments.

### Table 1: The detailed statistics of the used datasets.

|        | in-domain | out-of-domain | total |
|--------|-----------|---------------|-------|
| COCO[6]| VG[17]    | CC[34]        | SBU[29] |
| train  | 533K      | 5.1M          | 3.0M  | 869K | 9.5M |
| val    | 25K       | 106K          | 14K   | 10K  | 155K |

Multi-task Learning. Similar to [7], we adopt a multi-task learning objective to pretrain our model. Besides the proposed SKMLM and SKMRM tasks, we also include the image-text matching (ITM) task. Moreover, since the SKMLM and SKMRM tasks only focus on the key tokens included in the knowledge entry, we still retain the original random masking-based MLM and MRM tasks to guarantee a good coverage of the remaining tokens in the image and text.

5 EXPERIMENTS

We evaluate ROSITA on three V+L tasks and perform thorough comparative analysis to the state-of-the-art VLP methods on six datasets. Furthermore, we conduct comprehensive ablation experiments to explore its effectiveness in learning fine-grained semantic alignments.

5.1 Pretraining Setup

Datasets. Following the strategy in [6], we construct the pretraining dataset consisting of 9.5M train and 155K validation image-text pairs from four existing datasets, namely the COCO Captions [6], Visual Genome Captions [17], Conceptual Captions [34], and SBU Captions [29]. The four datasets are categorized into the in-domain and out-of-domain datasets based on whether they share the same images with the downstream tasks. The statistics of the pretraining dataset are shown in Table 1.

Implementation Details. For the input image-text pairs, we extract a fixed number of 36 region features from a pre-trained Faster R-CNN model [2] and adopt the BPE strategy to tokenize the sentence into a maximum of 50 words following [8]. Our ROSITA model adopts a 12-layer Transformer encoder architecture with 768 hidden units and 12 attention heads. The hyper-parameter $\alpha$ in Eq.(7) is set to 0.9. The masking probabilities in the original MRM and MLM tasks are set to 15% [45]. The model is initialized with the parameters from a pretrained BERT-base model [8], and then trained up to 40 epochs with a batch size of 512.

5.2 Downstream Tasks

After obtaining the pretrained ROSITA model, we finetune it on three downstream V+L tasks as follows.

Visual Question Answering (VQA) is a task that requires the model to answer natural language questions about an image. We adopt the widely used VQA v2 dataset [3, 10], which is manually built on the images from the MSCOCO dataset [24]. The dataset is split.

---

2We have made such an experiment that removes the MRM & MLM tasks. The resulting model reports slight performance drop (~0.3 points) on the downstream tasks.
Table 2: Results on downstream V+L tasks to compare with the state-of-the-art VLP methods. For a fair comparison, all the results are archived by the base models. Most of the models are trained on the in-domain+out-of-domain datasets, except for those models marked with † are trained on the out-of-domain datasets. IR and TR denote the image retrieval and text Retrieval, respectively. For the REC task, all the results are achieved based on the detected region features from images. Dark and light grey colors highlight the top and second best results on each evaluation metric.

| task | dataset | ViLBERT† | VLBERT† | Unicoder-VL | LXMERT | UNITER | ERNIE-ViL† | VILLA | OSCAR | ROSITA (ours) |
|------|---------|----------|----------|------------|--------|--------|-----------|-------|-------|--------------|
| VQA  | VQA v2  |          |          |            |        |        |           |       |       |              |
|      | test-dev | 70.55    | 71.16    | 72.42      | 72.70  | 72.62  | 73.59     | 73.16 | 73.44 | 73.91        |
|      | test-std | 70.92    | -        | -          | 72.54  | 72.91  | 72.85     | -     | -     |              |
| REC  | Ref-COCO |          |          |            |        |        |           |       |       |              |
|      | vald    | -        | -        | -          | -      | -      | 81.24     | -     | -     | 84.79        |
|      | testAd  | -        | -        | -          | -      | -      | 86.48     | -     | -     | 87.99        |
|      | testBd  | -        | -        | -          | 73.94  | -      | 74.48     | -     | -     | 78.28        |
| REC  | Ref-COCO+ |         |          |            |        |        |           |       |       |              |
|      | vald    | 72.34    | 71.60    | -          | 75.31  | 74.02  | 76.05     | -     | -     | 76.06        |
|      | testAd  | 78.52    | 77.72    | -          | 81.30  | 80.33  | 81.65     | -     | -     | 82.01        |
|      | testBd  | 62.61    | 60.99    | -          | 65.68  | 64.74  | 65.70     | -     | -     | 64.70        |
| REC  | Ref-COCOg |        |          |            |        |        |           |       |       |              |
|      | vald    | -        | -        | -          | 74.31  | -      | 75.90     | -     | -     | 78.23        |
|      | testd   | -        | -        | -          | 74.51  | -      | 75.93     | -     | -     | 78.25        |
| ITR  | IR-COCO |          |          |            |        |        |           |       |       |              |
|      | R@1     | -        | -        | 46.70      | -      | 50.33  | -         | 54.00 | -     | 54.40        |
|      | R@5     | -        | -        | 76.00      | -      | 78.52  | -         | 80.80 | -     | 80.92        |
|      | R@10    | -        | -        | 85.30      | -      | 87.16  | -         | 88.50 | -     | 88.60        |
| ITR  | TR-COCO |          |          |            |        |        |           |       |       |              |
|      | R@1     | -        | -        | 62.30      | -      | 64.40  | -         | 70.00 | -     | 71.26        |
|      | R@5     | -        | -        | 87.10      | -      | 87.40  | -         | 91.10 | -     | 91.62        |
|      | R@10    | -        | -        | 92.80      | -      | 93.08  | -         | 95.50 | -     | 95.58        |
| ITR  | IR-Flickr |       |          |            |        |        |           |       |       |              |
|      | R@1     | 58.20    | -        | 71.50      | -      | 72.52  | 74.44     | 74.74 | -     | 74.08        |
|      | R@5     | 84.90    | -        | 90.90      | -      | 92.36  | 92.72     | 92.86 | -     | 92.44        |
|      | R@10    | 91.52    | -        | 94.90      | -      | 96.08  | 95.94     | 95.82 | -     | 96.08        |
| ITR  | TR-Flickr |        |          |            |        |        |           |       |       |              |
|      | R@1     | -        | -        | 86.20      | -      | 85.90  | 86.70     | 86.60 | -     | 88.90        |
|      | R@5     | -        | -        | 96.30      | -      | 97.10  | 97.80     | 97.90 | -     | 98.10        |
|      | R@10    | -        | -        | 99.00      | -      | 98.80  | 99.00     | 99.20 | -     | 99.30        |

into train (83k images and 444k questions), validation (41k images and 214k questions), and test (81k images and 448k questions) sets. Following the strategy in [7], we feed the representation of the [CLS] token to a linear classifier to predict the corresponding answer from a vocabulary of size 3129 [49].

Referring Expression Comprehension (REC) is a task that requires to localize an image region referred to by a natural language query. We evaluate the performance on RefCOCO [16], RefCOCO+ [16] and RefCOCOg [28] datasets. All the three datasets are collected from COCO images [31]. RefCOCO and RefCOCO+ are split into four subsets, including train (120k queries), validation (11k queries), testA (6k queries about people), and testB (6k queries about objects), while RefCOCOg is split into three subsets, including train (81k queries), validation (5k queries), and test (10k queries). The representation for each image region is used to predict a ranking score and a refined bounding box.

Image-Text Retrieval (ITR) is a task that requires the model to calculate a similarity score between an image and a sentence and then perform cross-modal retrieval. We conduct experiments on the COCO Captions [6] and Flickr30K [44] datasets, respectively. Following the partition strategy by [15], the COCO dataset is split into 82k/5k/5k train/validation/test images, while the Flickr30K dataset is split into 29k/1k/1k train/validation/test images. Similar to [7], we use an offline hard sample mining strategy to obtain 128 negative samples per each positive sample, and use the representation of the [CLS] token to predict a matching score.

5.3 Main Results

We compare the proposed ROSITA model against existing state-of-the-art VLP methods. As shown in Table 2, ROSITA achieves the overall best performance on all downstream tasks, which verifies the effectiveness of the integrated cross- and intra-modal knowledge and the corresponding SKM strategy

It is worth noting that some methods like ViLBERT, LXMERT, and ERNIE-ViL adopt the two-stream architecture, which have much more parameters (ROSITA: 116M, VilBERT: 221M, LXMERT: 183M, ERNIE-ViL: 228M). Some methods like UNITER and VILLA use a larger number of image features (up to 100 regions), which has been verified to benefit the performance at the expense of much higher computational cost. In contrast, ROSITA uses a fixed number

3We have conduct such an experiment that pretrains ROSITA on the out-of-domain datasets only. The resulting model consistently outperforms the counterparts [27, 36, 45], verifying the generalization capability of our approach.
We run a number of ablations to investigate the reasons of ROSITA’s effectiveness. The results show in Table 3-4 and Figure 4-5 are discussed in detail below.

**Cross- and Intra-modal Knowledge.** In Table 3, we show the effects of the intra-modal knowledge and cross-modal knowledge based on the performance on the pretraining and downstream tasks. Taking the full ROSITA as the reference model (Line #1), we obtain the different variants by removing the cross-modal knowledge or the intra-modal knowledge. The variant without cross-modal knowledge (Line #2) indicates that the model is not aware of the anchor objects and the SKM strategy is performed only on a single modality using the intra-modal knowledge. In contrast, the variant without intra-modal knowledge (Line #3) indicates that the model is aware of the anchor objects but is not aware of the intra-modal contexts. Finally, by removing both the cross- and intra-modal knowledge, we obtain a baseline variant nearly identical to UNITER [7] (Line #4).

Given the pretrained models of the four variants above (i.e., without finetuning on downstream tasks), we evaluate their performance on three pretraining tasks. The ITM task examines the ability of semantic alignment between image-text pairs. From the results, we can see that both types of knowledge bring performance improvement to the ITM task (#4 vs. #3 and #2). Moreover, the two types of knowledge are complementary that their synergy brings 2.1 points improvement compared to the baseline model without any knowledge (#4 vs. #1). Although the ITM task is the most straightforward metric for semantic alignment, it only measures the coarse-grained alignments on the image-text level, thus cannot fully reveal the capability of ROSITA. As a complement, we resort to the SKMLM and SKMRM tasks to evaluate the fine-grained alignments on the region-word level. Compared with the baseline model in #4, the full ROSITA model improves the accuracies by 7.0 and 7.9 points on the SKMLM and SKMRM tasks, respectively.

Next, we report the performance of these variants on different downstream tasks. From the demonstrated results, we obtain similar observations to those on the pretraining tasks. The full ROSITA model consistently outperforms all the counterparts, verifying the effectiveness of the cross- and intra-modal knowledge.

**SKM Strategy.** After extracting knowledge entries from image-text pairs, we have two alternative masking strategies in SKM, i.e., independent probabilities and identical probability. All models are pretrained on the in-domain datasets and finetuned on the downstream tasks.

The effect of fine-grained semantic alignments across modalities can be inferred from the attention maps of the learned Transformer model [5]. We visualize the learned cross-modal attentions (i.e., region-to-words and word-to-regions attentions) from the pretrained UNITER [7] and our ROSITA models, as shown in Figure 4. Taking the image-text pair as inputs with exactly one token (a region or a word) being masked at a time, we pass the multimodal features through the pretrained model and extract the attention map from the last MSA block.

The region-to-words and word-to-regions attentions of the learned Transformer model [5]. We visualize the learned cross-modal attentions (i.e., region-to-words and word-to-regions attentions) from the pretrained UNITER [7] and our ROSITA models, as shown in Figure 4. Taking the image-text pair as inputs with exactly one token (a region or a word) being masked at a time, we pass the multimodal features through the pretrained model and extract the attention map from the last MSA block.

The region-to-words and word-to-regions attentions of the learned Transformer model [5]. We visualize the learned cross-modal attentions (i.e., region-to-words and word-to-regions attentions) from the pretrained UNITER [7] and our ROSITA models, as shown in Figure 4. Taking the image-text pair as inputs with exactly one token (a region or a word) being masked at a time, we pass the multimodal features through the pretrained model and extract the attention map from the last MSA block.

Table 4: Ablations of four ROSITA variants with two alternative masking strategy in SKM (i.e., independent probabilities and identical probability). All models are pretrained on the in-domain datasets and finetuned on the downstream tasks.

| masking prob.       | VQA2 (dev) | VQA2 (test) | RefCOCO (val) | IR-Flickr (test) | TR-Flickr (test) |
|---------------------|-----------|------------|--------------|-----------------|-----------------|
| independent         | 73.19     | 84.22      | 85.09        | 94.33           |                 |
| identical (p=15%)   | 72.93     | 84.36      | 93.63        |                 |                 |
| identical (p=30%)   | 72.93     | 84.36      | 93.63        |                 |                 |
| identical (p=45%)   | 72.75     | 82.96      | 83.75        | 93.33           |                 |

Our model has slight performance deviations compared with the original UNITER model since we use different visual features and pretraining hyper-parameters.

We perform element-wise addition over the attention maps from different heads followed by row-wise softmax normalization to obtain one aggregated attention map.
Figure 4: Visualizations of the learned cross-modal attentions (i.e., region-to-words attentions on the left and word-to-regions on the right) from UNITER [7] and ROSITA. Taking the image-text pair as inputs with exactly one region (or word) being masked at a time, we extract the attention map from the last MSA block of the pretrained model. The region-to-words (word-to-regions) attentions correspond to one specific row in the bottom-left (top-right) area of the attention map, respectively.

Figure 5: Visualizations of the region-to-words attentions (left) and word-to-regions attentions (right) from a pretrained ROSITA model with masking multiple regions (or words) at one time.

token correspond to one specific row in the bottom-left and top-right area of the attention map, respectively.

From the visualized cross-modal attentions, we can see that ROSITA learns significantly better semantic alignments than UNITER. ROSITA can precisely align the masked object to its reference object in the opposite modality while UNITER fails to establish such cross-modal alignments. For example, when the region of “ramp” is masked, ROSITA activates the word “ramp” precisely while UNITER obtains the largest attention value on the word “skate”. When another region of “person” is masked, ROSITA precisely activates the word “person” while UNITER still activates the incorrect word “skate”. Similar phenomena are observed in the opposite direction. ROSITA activates the accurate regions to the masked words while UNITER fails to do it.

To step further, we conduct a more challenging task as follows. We mask multiple regions (or words) at the same time to examine whether the semantic alignments can still be achieved. The visualized results in Figure 5 show that ROSITA works surprisingly well to establish accurate semantic alignment for each masked token. For example, when the regions of “bowl” and “bowl & carrots” are masked simultaneously, the region of “carrot” is precisely aligned to the word “carrots”, and the region of “bowl & carrots” is aligned to the two words “bowl” and “carrots” uniformly. In the opposite direction, when the words “man”, “candles”, and “cake” are masked at the same time, their corresponding regions are highlighted in the learned attentions, respectively.

6 CONCLUSION

In this paper, we present a new VLP method called ROSITA, which integrates the cross- and intra-modal knowledge in a unified scene graph to enhance the learning of cross-modal semantic alignment. We introduce a novel structural knowledge masking (SKM) strategy to perform masked language (region) modeling with respect to the knowledge entries extracted from the unified scene graph. Extensive ablations, comparative experiments, and comprehensive analysis show that ROSITA significantly outperforms existing state-of-the-art VLP approaches on three typical V+L tasks over six benchmark datasets. We hope our study will be helpful to inspire future research in the vision-and-language community and beyond.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0100603, and in part by National Natural Science Foundation of China under Grant 62072147, Grant 61836002, and Grant 62020106007.
Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In Proceedings of the 37th international ACM SIGIR Conference on Research & development in Information Retrieval. 395–404.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6281–6290.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. IEEE International Conference on Computer Vision (ICCV) (2017), 1839–1848.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems 29, 12 (2018), 5947–5959.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. DeVL bert: Learning Deconfounded Visio-Linguistic Representations. In Proceedings of the 28th ACM International Conference on Multimedia. 4373–4382.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 1441–1451.

Mingyang Zhou, Luowei Zhou, Shuchang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4155–4165.