Image-Text Embedding Learning via Visual and Textual Semantic Reasoning

Kunpeng Li, *Member, IEEE*, Yulun Zhang, *Member, IEEE*, Kai Li, *Member, IEEE*, Yuanyuan Li, *Student Member, IEEE*, and Yun Fu, *Fellow, IEEE*

Abstract—As a bridge between language and vision domains, cross-modal retrieval between images and texts is a hot research topic in recent years. It remains challenging because the current image representations usually lack semantic concepts in the corresponding sentence captions. To address this issue, we introduce an intuitive and interpretable model to learn a common embedding space for alignments between images and text descriptions. Specifically, our model first incorporates the semantic relationship information into visual and textual features by performing region or word relationship reasoning. Then it utilizes the gate and memory mechanism to perform global semantic reasoning on these relationship-enhanced features, select the discriminative information and gradually grow representations for the whole scene. Through the alignment learning, the learned visual representations capture key objects and semantic concepts of a scene as in the corresponding text caption. Experiments on MS-COCO [1] and Flickr30K [2] datasets validate that our method surpasses many recent state-of-the-arts with a clear margin. In addition to the effectiveness, our methods are also very efficient at the inference stage. Thanks to the effective overall representation learning with visual semantic reasoning, our methods can already achieve very strong performance by only relying on the simple inner-product to obtain similarity scores between images and captions. Experiments validate the proposed methods are more than 30-75 times faster than many recent methods with code public available. Instead of following the recent trend of using complex local matching strategies [3], [4], [5], [6] to pursue good performance while sacrificing efficiency, we show that the simple global matching strategy can still be very effective, efficient and achieve even better performance based on our framework.

Index Terms—Image-text retrieval, vision and language, cross-modal representation learning, graph neural networks, deep learning

1 Introduction

Vision and language are two important aspects of human intelligence to understand the real world. Many research explorations [3], [7], [8], [9] have been done in the computer vision as well as natural language processing areas to bridge these two modalities. As one of the fundamental topics in this field, image-text matching refers to measuring the visual-semantic similarity between an image and a text description. It has potential to benefit other vision-language tasks [10] and also can be applied to various applications, e.g., image search for given sentences or the retrieval of text descriptions from image queries.

Despite the encouraging progress achieved in this field, it is still challenging due to the huge visual semantic discrepancy. Intuitively, when people describe what they see in the picture using natural language, it can be observed that the descriptions will not only include the objects, salient stuff, but also will organize their interactions, relative positions and other high-level semantic concepts (such as "in mid-air" and "watching in the background" in the Fig. 1). During this process, it is crucial for humans to conduct visual semantic reasoning on objects and contexts in the scene. However, such kind of reasoning mechanism is missing in the existing visual-text matching systems. Most of them [7], [11], [12] represent concepts in an image by Convolutional Neural Network (CNN) features extracted by convolutions with a specific receptive field, which only performs local pixel-level analysis. Therefore, high-level semantic concepts cannot be well recognized by these methods. More recently, region-level features obtained from object detectors have been used to represent images [13]. Similarity between images and sentences can be calculated by aggregating from pairwise similarities among visual patches and words using complicated attention-weighted algorithms [3], [5]. Although these methods can capture some local semantic concepts within regions including multiple objects, they still lack the global reasoning mechanism, which allows information communication between regions farther away when generating the overall representations for images and sentences.

To address this issue, we propose Visual Semantic Reasoning Network (VSRN) to generate overall representations that capture key objects as well as their semantic relationships. Inspired by bottom-up attention mechanism [3], [13], we begin with salient region detection at stuff or object level in images, which is consistent with the human vision system [14]. Practically, we follow the protocol to implement the bottom-up attention module using Faster R-CNN [15]. As shown in Fig. 1, we then build up both semantic and spatial connections between these salient regions and utilize the

- *Kunpeng Li is with Facebook Reality Labs, Burlingame, CA 94010 USA. E-mail: kunpengli@ece.neu.edu.*
- *Yulun Zhang is with Computer Vision Lab, ETH Zürich, 8092 Zürich, Switzerland. E-mail: yulun100@gmail.com.*
- *Kai Li is with NEC Laboratories America Inc., Princeton, NJ 08540 USA. E-mail: kaili@ece.neu.edu.*
- *Yuanyuan Li and Yun Fu are with Northeastern University, Boston, MA 02115 USA. E-mail: {yuanyuanli, yunfu}@ece.neu.edu.*

Manuscript received 13 July 2021; revised 22 Dec. 2021; accepted 18 Jan. 2022. Date of publication 7 Feb. 2022; date of current version 5 Dec. 2022.
This work was supported in part by NSF IIS under Grant 1651902 and in part by U.S. Army Research Office under Grant W911NF-17-1-0367.
(Corresponding author: Yulun Zhang.)
Recommended for acceptance by K. Livescu.
Digital Object Identifier no. 10.1109/TPAMI.2022.3148470

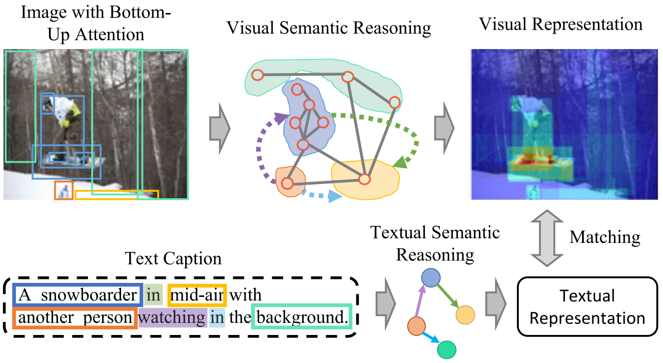


Image with Bottom-Up Attention
Visual Semantic Reasoning
Visual Representation
Matching
Text Caption
Textual Semantic Reasoning
A snowboarder in mid-air with another person watching in the background.
Textual Representation

Fig. 1. The proposed Visual Semantic Reasoning Network (VSRN++) performs reasoning on the image regions as well as words in the text caption to generate representations. Through the alignment learning, the learned visual representation captures key objects (boxes in the caption) and semantic concepts (highlight parts in the caption) of a scene as in the corresponding text caption.

power of Graph Convolutional Networks (GCN) [16] with residual structures to perform region relationship reasoning. Visual features enhanced with semantic and spatial relationship information can be obtained via this process.

When people describe the image as a whole, different regions and semantic relationships would have different levels of importance. Some of them are primary elements when photographers taking the pictures but some of them may be redundant considering the context of the scene. We aim to capture these important ones when generating the overall representation for the image. To achieve this goal, we propose to use the gate and memory mechanism [17] to perform global semantic reasoning on these relationship-enhanced features, select the discriminative information and gradually grow representation for the whole scene. This reasoning process is conducted on a graph topology and considers both local, global semantic correlations. The generated overall image representation should well capture the inner logic within the scene and pay more attention to key semantic concepts.

We also perform similar reasoning procedures on word vectors to obtain overall representations for sentences. The similarities between overall representations of images and sentences can be obtained by simple inner product. Finally, the whole alignment model is trained with joint optimization of Sinkhorn-based image-sentence matching and sentence generation, where we expect the learned visual representation should also have the ability to generate reasonable sentences that are similar to the ground-truth captions. Experiments on MS-COCO [1] and Flickr30K [2] datasets show that the proposed method surpasses many recent state-of-the-arts (SOTA) with a clear margin. In addition to the effectiveness, our method is also very efficient at the inference stage. Compared with recent popular methods that usually rely on complicated matching algorithms, e.g., SCAN [3], BFAN [4], CAMP [5], IMRAM [18], our model is more than 30-75 times faster on testing sets with different sizes. We validate that the classic and simple global matching strategy can still be very effective based on our visual semantic reasoning framework.

To investigate the interpretability of our model, we further design an interpretation method to analyze what has been learned. Correlations between the final image representation and each region feature are considered in this method. It well shows the model attention by tracing back to the individual patches about the significance. We find the learned image representation has high response at these regions that include key semantic concepts. Besides, we also visualize the top edges in the relationship reasoning layer to explain the behaviors of the model. We have interesting findings from these visualizations. Without using strong supervision such as scene graph or knowledge graph, the model can capture some reasonable semantic relationships when learning the alignments between images and captions.

To sum up, our main contributions are as follows:

- We propose an intuitive and interpretable model that conducts visual semantic reasoning on objects and contexts in the scene to learn a common embedding space for alignments between images and text descriptions.
- We incorporate the semantic relationship information into visual and textual representations by designing modules to perform region and word relationship reasoning.
- We propose to utilize the gate and memory mechanism to perform global semantic reasoning on the relationship-enhanced features, select the discriminative information and gradually grow representations for the whole scene.
- We conduct extensive experiments on MS-COCO [1] and Flickr30K [2] datasets to show that our model surpasses many recent methods with a clear margin.

This paper is an extended version of our previous work [19]. In particular: (a) We explore to incorporate a new region location relationship reasoning module into our visual reasoning framework, so that it encourages to learn better visual representations enhanced by spatial relationships and benefits the final matching performance. (b) We propose an extended framework so that it can perform semantic reasoning on the pre-trained textual features such as BERT. Experiments validate that this new module can help to generate better overall text representations, which improves the image-text matching results. (c) We investigate a new Sinkhorn-based image-text matching algorithm in our framework and validate its effectiveness. This is inspired by recent works in graph matching [20], [21] and optimal transport [22] areas. (d) We include considerable new experimental results and ablation studies to demonstrate the effectiveness of our method. We include more recently published works into comparisons and validate that our method outperforms many recent state-of-the-art methods in this area. (e) We validate the inference efficiency of our method, which is more than 30-70 times faster than several SOTA methods with code available. Instead of following the complex local matching algorithms to pursue good performance while sacrificing the efficiency, we show that the classic and simple global matching strategy can still be very effective based on our visual semantic reasoning framework. (f) We add additional detailed analyses and more quantitative visualizations in terms of model attention and edges in the relationship reasoning layer, which helps to interpret the behaviors of the model.

2 Related Work

Image-Text Matching. As a hot research topic to bridge the vision and language domains, the key issue of image-text matching is to measure the visual-semantic similarity between an image and a sentence. Our work is related to typical solutions proposed in this area that aim to learn a common embedding space, where image and sentence feature vectors are directly comparable. Frome *et al.* [23] use Skip-Gram and CNN to obtain image and text representations and construct a feature embedding framework accordingly. A ranking loss is then utilized to constrain the distance between the correctly matched image-sentence pair is smaller than that between the mismatched one in this embedding space. Kiros *et al.* [24] use a similar framework and adopt LSTM [25] instead of Skip-Gram for the learning of text representations. Vendrov *et al.* [26] design a new objective function that encourages the order structure of visual semantics can be preserved hierarchy. Yan *et al.* [27] adopt Canonical Correlation Analysis (CCA) to project the image and sentence with a projection which maximizes the correlation between them. To learn nonlinear projections, they cast CCA into a deep learning framework, which also improves the scalability of CCA to large training sets. Faghri *et al.* [7] propose an improved triplet loss function to focus more on hard negatives and achieve good improvement. There are other recent works that adopt image captioning technologies for the matching purpose. Fang *et al.* [28] first translate images into sentences by image captioning and then achieve alignments by comparing the similarity of generated sentences with ground-truth ones. Mao *et al.* [29] introduce a multimodal RNN structure to generate sentences from images in which the perplexity of generating a sentence is used as the similarity for alignments. However, their performance on measuring the image-sentence similarity is not very well due to the limitation of sentence generation quality. To address this issue, Gu *et al.* [12] further propose to learn better cross-view feature embedding by including generative objectives into the learning framework, which helps to improve matching performance by incorporating captioning. More recent methods [3], [18], [30] work on measuring similarities between images and captions by densely modeling correlations between image region features and word features. They have achieved very promising performance while sacrificing efficiency especially at the inference stage. Huang *et al.* explore a new few-shot setting for image-text matching [31] and show that the proposed model with cross-modal memory can achieve better caption retrieval results in both new and common settings. Our work follows the classic direction to learn a joint embedding space for sentences and images but with an emphasis on improving representation learning via reasoning.

We also notice that some recent works [32], [33], [34], [35], [36] explore pre-training using large-scale external image-text data pairs to improve the performance. They mainly employ self-attention mechanism found in Transformer models and set up training objectives similar to BERT [37] to learn cross-modal representations from a concatenated-sequence of visual region features and language token embeddings. However, exploring such cross-modal pre-training requires large amounts of annotated image-text pairs and huge computing resources, which are not easy to obtain in our case. We have included some of them [32], [33] into comparisons under the setting of using pre-trained BERT embedding to show the potential of our method. We leave the explorations of large-scale pre-training as future work.

Attention Mechanism. Our work is also closely related to bottom-up attention mechanism and recent explorations about interpretability of deep neural networks via attention visualization. Bottom-up attention [13], [14] refers to salient stuff/object detection, which can be analogized to the spontaneous bottom-up attention that is consistent with human vision system [13], [14], [38]. This observation has motivated several recent works in the image-text matching area. Karpathy *et al.* [39] adopt R-CNN to detect image regions and encode them to obtain object-level representations. The overall image-text similarity is then calculated by aggregating similarity scores for all region-word pairs. More recently, Huang *et al.* [40] train a multi-label CNN to classify each image region into multi-labels of objects and semantic relations, so that the improved image representation can capture semantic concepts within the local region. Lee *et al.* [3] present an attention algorithm to focus on key words and image regions when aggregating pairwise similarities among visual patches and words. Wang *et al.* [5] further study the cross-modal interactions between image regions and words and properly adopt an adaptive gating scheme to deal with mismatched pairs. We also follow bottom-up attention mechanism and start from object/region-level features of an image. To the best of our knowledge, our work is the first one that incorporates semantic and relation reasoning to learn overall representations for image-text matching.

Attention visualization is also one effective way to interpret the behaviors of neural networks. Cao *et al.* [41] propose a feedback method to capture the top-down neural attention. Gradient back propagation based methods [42], [43], [44], [45], [46], [47] are also very popular recently. They interpret the gradient of the prediction score of a particular class respecting to the inner layer features or original input image. Attention visualization is done by locating regions that are helpful for predicting a class. Following these explorations, we investigate the interpretability of our reasoning model with a new design of attention visualization. The proposed attention-based visualization strategy elegantly addresses a common problem for image-level similarity measure, which traces back to the individual patches about the significance.

Relational Reasoning Methods. Relational reasoning is initially introduced into the artificial intelligence community as symbolic methods [48]. They define relations between abstract symbols relying on the language of mathematics and logic, and then perform reasoning by deduction [49] etc. However, symbols need to be grounded [50] before such systems are practically useful. Modern approaches, such as path ranking algorithm [51], rely on statistical learning to extract useful patterns to perform relational reasoning on structured knowledge bases. As an active research area, graph-based methods [52] have been very popular in recent years and shown to be an efficient way of relation reasoning. CRFs [53] and random walk networks [54] are proposed based on the graph model for effective image segmentation. Recently, Graph Convolution Networks (GCN) [16] are proposed for semi-supervised classification. Yao *et al.* [55] train a visual relationship detection model on Visual Genome dataset [56] and use a GCN-based encoder to encode the detected relationship information into an image captioning framework. Yang *et al.* [57] utilize GCNs to incorporate the prior knowledge into a deep reinforcement learning framework,

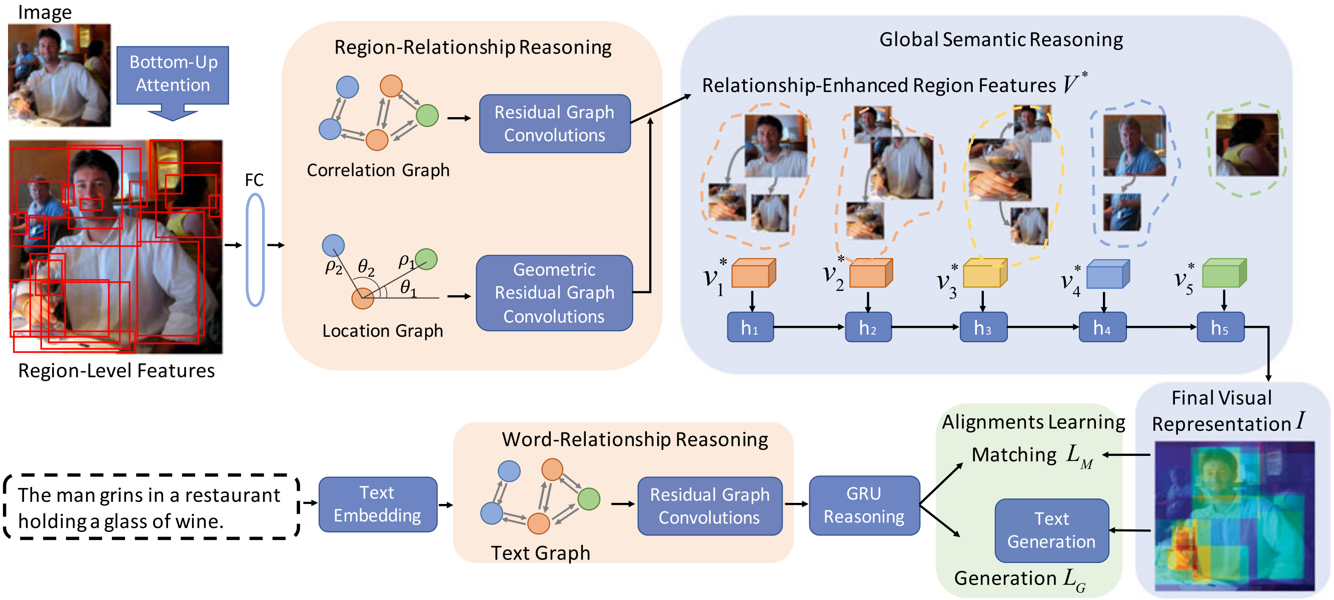


Image
Bottom-Up Attention
Region-Level Features
FC
Region-Relationship Reasoning
Correlation Graph
Residual Graph Convolutions
Location Graph
Geometric Residual Graph Convolutions
Global Semantic Reasoning
Relationship-Enhanced Region Features V^*
v_1^*
v_2^*
v_3^*
v_4^*
v_5^*
h1
h2
h3
h4
h5
Final Visual Representation I
Alignments Learning
Matching L_M
Text Generation
Generation L_G
Word-Relationship Reasoning
The man grins in a restaurant holding a glass of wine.
Text Embedding
Text Graph
Residual Graph Convolutions
GRU Reasoning

Fig. 2. An overview of the proposed Visual Semantic Reasoning Network. For the visual encoding part, starting from features of salient image regions obtained by the bottom-up attention model (Section 3.1), it first performs region relationship reasoning on these regions to generate features enhanced by both semantic and spatial location relationships (Section 3.2). Then the model applies gate and memory mechanisms to perform global semantic reasoning on the relationship enhanced features, select the discriminative information and gradually generate the representation for the whole image scene (Section 3.4). For the text encoding part, we design similar reasoning steps on the text caption with GCN and GRUs. This aims to improve the ability of capturing semantic context in the caption when generating the text embedding (Section 3.5). Finally, the whole model is trained with joint optimization of Sinkhorn-based image-text matching and sentence generation (Section 3.6). The attention of the visual representation (top right) is obtained by calculating correlations between the final image representation and each region feature (Section 4.5).

which improves semantic navigation in unseen scenes and towards novel objects. Other recent works [58], [59], [60] apply Graph Learning framework for visual commonsense reasoning and aim to predict correct answers for the given question and meanwhile provide convincing reasoning paths. They either focus on dynamically reorganize the visual neuron connectivity that is contextualized by the meaning of questions and answers or develop the vision-to-answer heterogeneous graph.

We also adopt the reasoning power of graph convolutions to obtain image region features and word features enhanced with semantic relationship. But we do not need extra database to build the relation graph (e.g. [55] needs to train the relationship detection model on Visual Genome). Beyond this, we propose GCN-GRU structure to further perform global semantic reasoning on the relationship-enhanced features obtained by GCN, so that the final global representation can capture key objects and semantic concepts of a scene. Such structure shares some similar high-level ideas with self-attention mechanisms in Transformers. They both consider correlations between image regions when learning enhanced representations. Differently, our GCNs does not have the attentional behavior (softmax in self-attention) to constrain the descriptions of semantic correlations between image patches. The proposed model also incorporates location relationships into visual representation learning, while we have not seen such abilities or usages for Transformers.

3 Learning Alignments With Visual Semantic Reasoning

We describe the detailed structure of the Visual Semantic Reasoning Network (VSRN) for image-text matching as shown in Fig. 2. Our goal is to infer the similarity between a full sentence and a whole image by mapping image regions and the text descriptions into a common embedding space. For the image part, we begin with image regions and their features generated by the bottom-up attention model [13] (Section 3.1). VSRN first builds up connections between these image regions and do reasoning using Graph Convolutional Networks (GCN) to generate features with semantic relationship information (Section 3.2). Then, we do global semantic reasoning on these relationship-enhanced features to select the discriminative information and filter out unimportant one to generate the final representation for the whole image (Section 3.4). For the text caption part, in addition to learning an embedding for the sentence using RNNs, we also explore reasoning for textual representation learning to improve the ability of capturing the semantic context (VSRN++, Section 3.5). Finally, the whole model is trained with joint optimization of matching and generation objectives (Section 3.6).

3.1 Image Representation by Bottom-Up Attention

Taking the advantage of bottom-up attention [13], each image can be represented by a set of features $V = \{v_1, \ldots, v_k\}, v_i \in \mathbb{R}^D$, such that each feature v_i encodes an object or a salient region in this image. Following [3], [13], we implement the bottom-up attention with a Faster R-CNN [15] model using ResNet-101 [61] as the backbone. It is pre-trained on the Visual Genomes dataset [56] by [13]. The model is trained to predict instance classes and attribute classes instead of the object classes, so that it can help learn feature representations with rich semantic meaning. Specifically, instance classes include objects and salient stuff which is hard to recognize. For example, attributes like "furry" and stuff like "building", "grass" and "sky". The model's final output is used and non-maximum suppression for each class is operated with an IoU threshold of 0.7. We then set a confidence threshold of 0.3 and select all image regions where any class detection probability is larger than this threshold. The top 36 ROIs with the highest class detection confidence scores are selected. All these thresholds are set as same as [3], [13]. For each selected region i, we extract features after the average pooling layer, resulting in f_i with 2048 dimensions. A fully-connect layer is then applied to transform f_i to a D-dimensional embedding using the following equation:

$$v_i = W_f f_i + b_f. \tag{1}$$

Then $V = \{v_1, \ldots, v_k\}, v_i \in \mathbb{R}^D$ is constructed to represent each image, where v_i encodes an object or salient region in this image.

3.2 Region Semantic Relationship Reasoning

Inspired by recent advances in deep learning based visual reasoning [62], [63], [64], we build up a region relationship reasoning model (RRR) to enhance the region-based representation by considering the semantic correlation between image regions. Specifically, we measure the pairwise affinity between image regions in an embedding space to construct their relationship using Eq. (2).

$$R(v_i, v_j) = \varphi(v_i)^T \phi(v_j), \tag{2}$$

where $\varphi(v_i) = W_\varphi v_i$ and $\phi(v_j) = W_\phi v_j$ are two embeddings. The weight parameters W_φ and W_ϕ can be learned via back propagation.

Then a fully-connected relationship graph $G_r = (V, E)$, where V is the set of detected regions and edge set E is described by the affinity matrix R. R is obtained by calculating the affinity edge of each pair of regions using Eq. (2). That means there will be an edge with high affinity score connecting two image regions if they have strong semantic relationships and are highly correlated.

We apply the Graph Convolutional Networks (GCN) [16] to perform reasoning on this fully-connected graph. Response of each node is computed based on its neighbors defined by the graph relations. We add residual connections to the original GCN as follows:

$$V^\star = (RVW_g)W_r + V, \tag{3}$$

where W_g is the weight matrix of the GCN layer with dimension of $D \times D$. W_r is the weight matrix of the residual structure. R is the affinity matrix with shape of $k \times k$. We follow the routine to row-wise normalize the affinity matrix R. The output $V^\star = \{v_1^\star, \ldots, v_k^\star\}, v_i^\star \in \mathbb{R}^D$ is the semantic relationship enhanced representation for image region nodes.

3.3 Region Location Relationship Reasoning

In addition to the reasoning based on correlations between region representation, we also propose to conduct region location relationship reasoning (LRR) to capture the spatial relationships among these object patches. Inspired by recent graph machine learning methods on the geometric data [65], [66], we adopt a geometric CNN model on the location graph in the image. The location relationships between object patches are obtained by a pairwise pseudo-coordinate function $p(i, j)$ that defines the relative location of image patch v_j towards the image patch v_i. Specifically, we utilize the geodesic polar coordinates [66] (ϕ, θ) as the return from $p(i, j)$, which provides both distance and orientation information between two image patches.

To describe the influence of the neighborhood nodes, we follow [66], [67], [68] to process the polar coordinates with a set of Gaussian kernels $\kappa = \{\kappa_1(\cdot), \kappa_2(\cdot), \ldots, \kappa_M(\cdot)\}$ whose means and covariances are learnable parameters. We refine the original geometric CNN model by adding residual structures and the final operation on node v_i is as follows:

$$v_i^\# = \left[\bigcup_{m=1}^{M} W_m^\# \sum_{j \in \mathcal{N}(i)} \kappa_m(p(i, j)) v_j \right] W_r^\# + v_i, \tag{4}$$

where $\mathcal{N}(i)$ is the set of neighborhood nodes for current node i. $W_m^\# \in \mathbb{R}^{\frac{D}{M} \times D}$ is the weight matrix of the convolution operation. $\bigcup$ is the concatenation operation on the results of M kernels κ. $W_r^\#$ is the weight matrix of the residual structure.

We conduct the geometric CNN operations on each image patch and obtain the location relationship enhanced representation for image regions as $V^\# = \{v_1^\#, \ldots, v_k^\#\}, v_i^\# \in \mathbb{R}^D$. Then we combine $V^\#$ with the semantic relationship enhanced feature $V^\star$ (obtained by Eq. (3)) with a add operation followed by L_2 normalization to obtain final V^*. V^* is the region representations enhanced by both semantic and spatial location relationships.

3.4 Global Semantic Reasoning

Based on region features with relationship information, we further do global semantic reasoning to select the discriminative information and filter out unimportant one to obtain the final representation for the whole image. Specifically, we perform this reasoning process by putting the sequence of graph enhanced region features $V^* = \{v_1^*, \ldots, v_k^*\}, v_i^* \in \mathbb{R}^D$ into GRUs [17]. The description of the whole scene will gradually grow and update in the memory cell (hidden state) m_i during this process described as follows: At each step i, an update gate z_i analyzes the current input region feature v_i^* and the description of the whole scene at last step m_{i-1} to decide how much the unit updates its memory cell. The update gate is calculated by:

$$z_i = \sigma_z(W_z v_i^* + U_z m_{i-1} + b_z), \tag{5}$$

The newly added content helping grow the description of the whole scene is computed as follows:

$$\tilde{m}_i = \sigma_m(W_m v_i^* + U_m(r_i \circ m_{i-1}) + b_m), \tag{6}$$

where r_i is the reset gate that decides what content to forget based on the correlations between v_i^* and m_{i-1} and it is computed similarly to the update gate. σ_z and σ_m are activation functions, W_z, U_z, W_m, U_m are weights and b_z, b_m are biases. $\circ$ is an element-wise multiplication. Then the description of the whole scene m_i at the current step is obtained by

$$m_i = (1 - z_i) \circ m_{i-1} + z_i \circ \tilde{m}_i, \tag{7}$$

Since each v_i^* includes global relationship information, updates of m_i are actually based on reasoning on a graph topology, which considers both local region and global semantic correlations. We take the last memory cell m_k as the final representation I for the whole image, where k is the length of V^*.

3.5 Text Representation

In order to build connections between language and vision domains, we also need to encode the text caption to the same D-dimensional semantic vector space $C \in \mathbb{R}^D$, which shares the same dimension with the image representation I.

To achieve this, the most straightforward way is to map each word w_i in the given text caption into the vector space through an embedding matrix as follows:

$$t_i = W_t w_i + b_t, \tag{8}$$

where $T = \{t_1, \ldots, t_n\}, t_i \in \mathbb{R}^{D'}$ is obtained to represent each text caption, where t_i encodes a word in this sentence.

However, this approach does not take any semantic context in the text caption into consideration when generate the text embedding. Therefore, we then apply GRUs [17] to process the sequence of word embedding $T = \{t_1, \ldots, t_n\}$ and generate the whole representation $C \in \mathbb{R}^D$ of the text caption. This is how we encode the text caption in our conference version.

To further improve the ability of capturing the semantic context, we follow [69] to take use of Bidirectional Encoder Representations from Transformers (BERT) [37] to encode words in captions. BERT is pre-trained from unlabeled text to obtain deep bidirectional representations by jointly conditioning on both right and left textual context in all network layers. For each given text caption, we adopt BERT to encode it with word representations $T = \{t_1, \ldots, t_n\}, t_i \in \mathbb{R}^{D'}$, $D' = 768$. Different from common practice of fine-tuning the BERT end-to-end towards the target tasks, we do not perform fine-tuning because it will take too much GPU memory for the image-text matching task where the batch size is always large. Instead, we follow [69] to extract feature T from the pre-trained BERT model. We then build up textual reasoning model to learn representation capturing semantic context for alignment between text and image.

Although the BERT feature is known to have good ability of capturing long-term complex relations of words in the sentence with self-attention, we find it is not ideal if only relying on the pre-trained BERT feature. In the our case, we keep the same visual encoding part in VSRN and adopt a fully connected layer followed with a mean-pooling operation to map the BERT feature to the same dimension as the visual representation, so that the visual and textual feature can be directly comparable. Experimental results ("BERT + FC Mean-pool" in Table 5) show limited performance on the matching task. Therefore, we extend our visual encoding framework to the text encoding part to learn relationship enhanced text representations.

Specifically, to consider the semantic correlation between words in the given caption, we measure the pairwise affinity between words in an embedding space to construct their relationship similar to Eq. (2) as: $R'(t_i, t_j) = \varphi'(t_i)^T \phi'(t_j)$, where $\phi'(\cdot)$ and $\varphi'(\cdot)$ are two embedding functions with learnable weight parameters. Similar to the visual encoding mode, we also build up a fully-connected relationship graph $G'_r = (T, E')$, where T is the word representation set and edge set E is represented by the affinity matrix R'. We then adopt the Graph Convolutional Networks with residual connections to perform word-relationship reasoning on this fully-connected graph:

$$T^* = (R' T W'_g) W'_r + T, \tag{9}$$

where W'_g is the weight matrix of the GCN layer, W'_r is the weight matrix of the residual structure. R' is the affinity matrix. The output $T^* = \{t_1^*, \ldots, t_k^*\}, t_i^* \in \mathbb{R}^{D'}$ is the relationship enhanced representation for words in the text caption.

With the relationship enhanced representation for words, we then adopt GRUs [17] to process the sequence of word embedding $T^* = \{t_1^*, \ldots, t_k^*\}$ and generate the whole representation $C \in \mathbb{R}^D$ of the text caption. The motivation of applying GRUs here is the same as Global Semantic Reasoning in the visual encoding part. This version of our model is noted as VSRN++.

3.6 Learning Alignments by Joint Optimization of Sinkhorn-Based Matching and Generation

To learn an alignment model that bridges the language and vision domains, we jointly optimize both the matching between caption representation C and image representation I as well as caption generation from I.

For the matching part, we first obtain the initial correspondence score matrix $\mathcal{S}^{(0)}$ between two sets of image embedding and text embedding by calculating the similarity between each pair of image-text embedding vectors by $\mathcal{S}^{(0)}_{i,j} = f_s(I_i, C_j)$, where we use inner product as $f_s(\cdot)$ in our experiments, $\mathcal{S}^{(0)} \in \mathbb{R}^{B \times B}$ and B is the size of the image and text embedding set (a batch during the training). Inspired by recent works in graph matching [20], [21] and optimal transport [22], we apply Sinkhorn Normalization [70], [71], [72] for stability. Sinkhorn algorithm is an efficient yet simple approximate solution for generalized linear assignment. It is a differentiable version of the Hungarian algorithm [73], classically used for bipartite matching. Specifically, Sinkhorn normalization is used to convert $\mathcal{S}^{(0)}$ to a doubly-stochastic matrix $\mathcal{S}^{(0)}$ by repeatedly normalizing rows and columns. The row and column normalization processes are defined as:

$$\mathcal{S}^{(k)'} = \mathcal{S}^{(k-1)} \oslash (\mathcal{S}^{(k-1)} \mathbf{1}\mathbf{1}^\top), \tag{10}$$

$$\mathcal{S}^{(k)} = \mathcal{S}^{(k)'} \oslash (\mathbf{1}\mathbf{1}^\top \mathcal{S}^{(k)'}), \tag{11}$$

where $\oslash$ is the Hadamard (elementwise) division and $\mathbf{1}$ is a vector of ones. The doubly-stochastic matrix S obtained by Sinkhorn normalization is treated as our model's matching prediction. $S = \mathrm{S}inkhorn(\mathcal{S}^{(0)})$. In our conference version (VSRN), we directly use $S^{(0)}$ as S without exploring the Sinkhorn-based matching. Finally, we adopt a hinge-based triplet ranking loss [3], [7], [39] with emphasis on hard negatives [7] for the matching objective. We define the loss as:

$$L_M = [\alpha - S(I,C) + S(I,\hat{C})]_+ \\ + [\alpha - S(I,C) + S(\hat{I},C)]_+, \qquad (12)$$

where α serves as a margin parameter. $[x]_+ \equiv \max(x,0)$. This hinge loss comprises two terms, one with I and one with C as queries. $S(\cdot)$ is the similarity score in the corresponding position in S. $\hat{I} = \arg\max_{j \neq I} S(j,C)$ and $\hat{C} = \arg\max_{d \neq C} S(I,d)$ are the hardest negatives for a positive pair (I, T). For computational efficiency, instead of finding the hardest negatives in the entire training set, we find them within each mini-batch.

For the generation objective, our motivation is that the learned visual representations should also have the ability to generate sentences that are close to the ground-truth captions. Specifically, we adopt a sequence-to-sequence model with attention mechanism [74] on V^* sequence to achieve this goal. This model incorporates a stacked LSTM and includes a encoding and decoding stage. Given the input sequence V^*, the LSTM computes a sequence of hidden states. During decoding it defines a distribution over the output sequence. The model maximizes the log-likelihood of the predicted output sentence with the following loss function:

$$L_G = -\sum_{t=1}^{l} \log p(y_t | y_{t-1}, V^*; \theta), \qquad (13)$$

where l is the length of output word sequence $Y = (y_1, \ldots, y_l)$. θ is the parameter of the sequence to sequence model.

Our final loss function is defined as follows to perform joint optimization of the two objectives.

$$L = L_M + L_G. \qquad (14)$$

Please note the generation part is only used to provide regularization on the representation learning of V^* at the training stage. It is not used and does not bring extra computation cost at the inference stage. As for the training stage, this part takes around 2 seconds time cost for each iteration with batch size of 128 in our experiments. The result 9 seconds/iteration in total for the full model is still very time efficient.

4 Experiments

To evaluate the proposed Visual Semantic Reasoning Network, we conduct experiments in terms of image retrieval (sentence query) and sentence retrieval (image query) on two standard benchmarks. Recent state-of-the-art methods on the image-text matching task are included for comparisons. We also design ablation studies to investigate the effectiveness of each model component.

4.1 Datasets and Protocols

We evaluate our method on the Flickr30K dataset [2] and the Microsoft COCO dataset [1]. Flickr30K includes 31783 images that are obtained from the Flickr website. Each image is corresponding to 5 text captions from human annotations. We use the standard training, validation and testing splits [39], which include 28,000 images, 1000 images and 1000 images respectively. MS-COCO contains 123,287 images, where each image is accompanied with 5 human annotated text descriptions. We follow the splits of [3], [7], [12], [39] for MS-COCO dataset, which include 113,287 images for training, 5,000 images for validation and 5000 images for testing. Each image corresponds to 5 text captions. There are two kinds of evaluation protocols. One is MS-COCO 5-folder 1K test set, where the final performance is obtained by averaging the results from 5 folds of 1K test images. Another is MS-COCO 5K test set, which means the testing is performed on the full 5K test images. Following common practice in information retrieval, we measure the matching performance by recall at K (R@K). This is defined as the fraction of queries for which the correct item is retrieved in the closest K points to the query.

4.2 Implementation Details

We follow the same procedure as [3], [13] to set up hyper parameters, backbone network, training dataset and other details of visual bottom-up attention model. The order of regions for GRU-based global semantic reasoning (Section 3.4) is determined by the descending order of their class detection confidence scores that are generated by the bottom-up attention detector. We set the dimension of the joint embedding space D as 2048 and the number of kernels $M = 8$ in the Location Relationship Reasoning module. The word embedding size is set as 300 for VSRN when using Eq. (8) to map words in captions. For VSRN++, we use 12-layer BERT model with hidden size of 768. Interactions of Sinkhorn normalization is set as 20.

For the training of VSRN, Adam optimizer [83] is adopted to train the model with 30 epochs. The learning rate is set as 0.0002 for the first 15 epochs, and then decreased to 0.00002 for the rest 15 epochs. For the training of VSRN++, we find the model is easier to converge, therefore, set the total epochs as 16 and lower the learning rate to 0.00002 after the training of first 8 epochs. The margin α in Eq. (12) is set as 0.2 and size of mini-batch is set as 128 in all our experiments. For model selection during training, we follow the same protocol with VSE++ [7] and SCAN [3]. We chose the snapshot of the model that has the best performance on the validation set to avoid over-fitting. The best performance is defined according to the sum of recalls on the validation set.

4.3 Comparisons With Several Recent State-of-the-Arts

Results on Flickr30K Dataset. In Table 1, we show results of VSRN and VSRN++ as well as comparisons with several recent state-of-the-art methods on Flickr30K test set. The model setting information is also listed in the table such as network backbones used for visual feature extraction, e.g., AlexNet, R-CNN object detector, VGG, ResNet and Faster R-CNN. We follow the same strategy with SCAN [3] by averaging predication scores from two trained models for fair comparison. The difference is that we do not design different configurations or versions of the model, we just simply train two models with the same configuration separately. We experimentally find this can help our method to achieve better results compared with the single model.

TABLE 1
Quantitative Results of the Image-to-Text (Caption) Retrieval and Text-to-Image (Image) Retrieval on Flickr30K Test set in Terms of Recall@K (R@K)

Methods	Visual Feature	Caption Retrieval			Image Retrieval			rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
RRF$_{\text{(ICCV'17)}}$ [75]	ResNet-101	47.6	77.4	87.1	35.4	68.3	79.9	396.0
VSE++$_{\text{(BMVC'18)}}$ [7]	ResNet-101	52.9	79.1	87.2	39.6	69.6	79.5	408.0
SCO$_{\text{(CVPR'18)}}$ [40]	ResNet-101	55.5	82.0	89.3	41.1	70.5	80.1	418.2
SCO$_{\text{(TPAMI'18)}}$ [76]	ResNet-101	58.0	84.5	90.5	43.9	72.9	81.6	431.4
(Pre-training With Extra Image-text Data Pairs)								
ViLBERT$_{\text{(NeurIPS'19)}}$ [32]	Faster R-CNN, ResNet-101	-	-	-	58.2	84.9	91.5	-
Unicoder-VL$_{\text{(AAAI'20)}}$ [33]	Faster R-CNN, ResNet-101	73.0	89.0	94.1	57.8	82.2	88.9	485.0
UNITER-large$_{\text{(ECCV'20)}}$ [35]	Faster R-CNN, ResNet-101	86.9	98.1	**99.2**	75.5	94.0	96.6	550.3
LightningDOT$_{\text{(NAACL'21)}}$ [77]	Faster R-CNN, ResNet-101	87.2	**98.3**	99.0	75.6	94.0	96.5	550.6
VILLA$_{\text{(NeurIPS'20)}}$ [78]	Faster R-CNN, ResNet-101	**87.9**	97.5	98.8	**76.3**	**94.2**	**96.8**	**551.5**
(Dense Local and Global Matching, without Pre-training)								
SCAN$_{\text{(ECCV'18)}}$ [3]	Faster R-CNN, ResNet-101	67.4	90.3	95.8	48.6	77.7	85.2	465.0
BFAN$_{\text{(ACM MM'19)}}$ [4]	Faster R-CNN, ResNet-101	68.1	91.4	95.9	50.8	78.4	85.8	470.4
CAMP$_{\text{(ICCV'19)}}$ [5]	Faster R-CNN, ResNet-101	68.1	89.7	95.2	51.5	77.1	85.3	466.8
MPL$_{\text{(CVPR'20)}}$ [79]	Faster R-CNN, ResNet-101	69.4	89.9	95.4	47.5	75.5	83.1	460.8
KASCE$_{\text{(IJCAI'19)}}$ [80]$*$	Faster R-CNN, ResNet-101	71.8	90.8	94.8	49.3	76.4	85.6	468.7
GOT$_{\text{(ICML'20)}}$ [81]	Faster R-CNN, ResNet-101	70.9	92.8	95.5	50.7	78.7	86.2	474.8
CAAN$_{\text{(CVPR'20)}}$ [6]	Faster R-CNN, ResNet-101	70.1	91.6	**97.2**	52.8	79.0	**87.9**	478.6
IMRAM$_{\text{(CVPR'20)}}$ [18]	Faster R-CNN, ResNet-101	74.1	**93.0**	96.6	53.9	79.4	87.2	484.2
MMCA$_{\text{(CVPR'20)}}$ [69]	Faster R-CNN, ResNet-101	**74.2**	92.8	96.4	**54.8**	**81.4**	87.8	**487.4**
(Only Global Matching, without Pre-training)								
ViLBERT$_{\text{(NeurIPS'19)}}$ [32]	Faster R-CNN, ResNet-101	-	-	-	45.5	76.8	85.0	-
GVSE$_{\text{(AAAI'19)}}$ [82]	Faster R-CNN, ResNet-101	68.5	90.9	95.5	50.6	79.8	87.6	472.8
VSRN (ours)	Faster R-CNN, ResNet-101	71.3	90.6	96.0	54.7	81.8	88.2	482.4
VSRN++ (ours)	Faster R-CNN, ResNet-101	**79.2**	**94.6**	**97.5**	**60.6**	**85.6**	**91.4**	**508.9**

$*$ *indicates using human annotated scene graphs.*

From results, we can see the proposed VSRN and VSRN++ outperform many recent state-of-the-art methods with a clear margin, especially for Recall@1. When compared with methods that build upon the same visual feature (Faster R-CNN with ResNet-101), our VSRN outperforms SCAN [3] by 5.8% on caption retrieval (R@1) and 12.6% on image retrieval(R@1) relatively, VSRN++ further achieves 17.5% on caption retrieval (R@1) and 24.7% on image retrieval(R@1) relative improvements upon SCAN. SCAN aims to find out all latent alignments between possible regions and words pairs. When inferring the image-text similarity, it further adopts an attention model to focus on key alignments. SCAN mainly focuses on local pairwise matching between regions and words. In contrast, the proposed VSRN and VSRN++ perform reasoning on region features and generate a global scene representation that captures key objects and semantic concepts for each image. This representation can be better aligned with the corresponding text caption. CAMP [5], CAAN [6] and IMRAM [18] are more recent methods that build upon SCAN. They also mainly considers local alignments between image regions and caption words with message passing across modalities. They are still different from our explorations of generating whole representations for the image scene.

Besides, as discussed in the related work, we also include several recent works that explore transformer structures and pre-training using large-scale external image-text data pairs e.g., ViLBERT [32], Unicoder-VL [33] etc. We find that without relying on pre-trianing with extra image-text data pairs, our models shows strength compared with Transformer-based models such as ViLBERT [32] and MMCA [69]. Transformer models show more advantages especially when conducting pre-training with a large amount of data available. In addition to these works, we also find there is a new few-shot setting for image-text matching [31] and notice the model proposed for this new setting can achieve better caption retrieval results by using the cross-modal memory.

Results on MS-COCO Dataset. We report results on MS-COCO 5-fold 1K test set as shown in Table 2, where the proposed VSRN and VSRN++ perform better than several recent state-of-the-art methods. Following the common protocol [3], [7], [40], results on MS-COCO 5-fold 1K test set are actually obtained by averaging over 5 folds of 1K test images. When compared with the most popular method SCAN [3], our VSRN outperforms it by 4.8% on caption retrieval (R@1) and 6.8% on image retrieval (R@1) relatively and VSRN++ further achieves 7.2% on caption retrieval (R@1) and 9.7% relative improvement upon SCAN. In Table 3, we also show results on MS-COCO 5K test set, which are obtained by testing on the full 5K test images and their corresponding captions. Among all methods, the proposed VSRN and VSRN++ still achieve the best performance, which is consistent with results on Flickr dataset and again demonstrates the effectiveness of learning the global image-text representations by visual-semantic reasoning.

Inference Efficiency Analysis. In addition to the accuracy of caption or image retrieval, we also argue that the efficiency at the inference stage are important when evaluating the model's performance. This is crucial especially when the model is used

TABLE 2
Quantitative Evaluation Results of the Image-to-Text (Caption) Retrieval and Text-to-Image (Image) Retrieval on MS-COCO 5-Fold 1K Test set in Terms of Recall@K (R@K)

Methods	Visual Feature	Caption Retrieval			Image Retrieval			rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
RRF$_{\text{(ICCV'17)}}$ [75]	ResNet-101	56.4	85.3	91.5	43.9	78.1	88.6	443.4
VSE++$_{\text{(BMVC'18)}}$ [7]	ResNet-101	64.6	89.1	95.7	52.0	83.1	92.0	476.4
GXN$_{\text{(CVPR'18)}}$ [12]	ResNet-101	68.5	-	97.9	56.6	-	94.5	-
SCO$_{\text{(CVPR'18)}}$ [40]	ResNet-101	69.9	92.9	97.5	56.7	87.5	94.8	499.2
SCO$_{\text{(TPAMI'18)}}$ [76]	ResNet-101	71.3	93.8	98.0	58.2	88.8	95.3	505.2
(Pre-training With Extra Image-text Data Pairs)								
Unicoder-VL$_{\text{(AAAI'20)}}$ [33]	Faster R-CNN, ResNet-101	75.1	94.3	97.8	63.9	91.6	96.5	519.2
(Dense Local and Global Matching, without Pre-training)								
SCAN$_{\text{(ECCV'18)}}$ [3]	Faster R-CNN, ResNet-101	72.7	94.8	98.4	58.8	88.4	94.8	508.2
CAMP$_{\text{(ICCV'19)}}$ [5]	Faster R-CNN, ResNet-101	72.3	94.8	98.3	58.5	87.9	95.0	506.8
BFAN$_{\text{(ACMMM'19)}}$ [4]	Faster R-CNN, ResNet-101	74.9	95.2	-	59.4	88.4	–	–
KASCE$_{\text{(IJCAI'19)}}$ [80]$*$	Faster R-CNN, ResNet-101	76.6	**96.3**	**99.2**	61.4	88.9	95.1	**517.5**
MPL$_{\text{(CVPR'20)}}$ [79]	Faster R-CNN, ResNet-101	71.1	93.7	98.2	56.8	86.7	93.0	499.5
MMCA$_{\text{(CVPR'20)}}$ [69]	Faster R-CNN, ResNet-101	74.8	95.6	97.7	61.6	**89.8**	**95.2**	514.7
CAAN$_{\text{(CVPR'20)}}$ [6]	Faster R-CNN, ResNet-101	75.5	95.4	98.5	61.3	89.7	95.2	515.6
IMRAM$_{\text{(CVPR'20)}}$ [18]	Faster R-CNN, ResNet-101	**76.7**	95.6	98.5	**61.7**	89.1	95.0	516.6
(Only Global Matching, without Pre-training)								
GVSE$_{\text{(AAAI'19)}}$ [82]	Faster R-CNN, ResNet-101	72.2	94.1	98.1	60.5	89.4	95.8	510.1
VSRN (ours)	Faster R-CNN, ResNet-101	76.2	94.8	98.2	62.8	89.7	95.1	516.8
VSRN++ (ours)	Faster R-CNN, ResNet-101	**77.9**	**96.0**	**98.5**	**64.1**	**91.0**	**96.1**	**523.6**

$*$ *indicates using human annotated scene graphs.*

in search engines for a large-scale database towards image or text queries. However, recent state-of-the-art methods [3], [4], [5], [6], [18] usually rely on complex local matching strategies. For example, in the representative work SCAN [3], the final matching is based on aggregation from pairwise similarities among visual patches and words (densely matching between two sets of vectors) using complicated attention-weighted algorithms. Although this kind of complex matching algorithm

TABLE 3
Quantitative Evaluation Results of the Image-to-Text (Caption) Retrieval and Text-to-Image (Image) Retrieval on MS-COCO 5K Test set in Terms of Recall@K (R@K)

Methods	Visual Feature	Caption Retrieval			Image Retrieval			rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++$_{\text{(BMVC'18)}}$ [7]	ResNet-101	41.3	69.2	81.2	30.3	59.1	72.4	353.4
GXN$_{\text{(CVPR'18)}}$ [12]	ResNet-101	42.0	-	84.7	31.7	-	74.6	-
SCO$_{\text{(CVPR'18)}}$ [40]	ResNet-101	42.8	72.3	83.0	33.1	62.9	75.5	369.6
SCO$_{\text{(TPAMI'18)}}$ [76]	ResNet-101	45.7	76.0	86.4	36.8	67.0	78.8	390.6
((Pre-training With Extra Image-text Data Pairs)								
Unicoder-VL$_{\text{(AAAI'20)}}$ [33]	Faster R-CNN, ResNet-101	62.3	87.1	92.8	46.7	76.0	85.3	450.2
UNITER-large$_{\text{(ECCV'20)}}$ [35]	Faster R-CNN, ResNet-101	65.7	88.6	93.8	52.9	79.9	88.0	468.9
OSCAR$_{\text{(ECCV'20)}}$ [36]	Faster R-CNN, ResNet-101	73.5	92.2	96.0	**57.5**	**82.8**	89.8	491.8
LightningDOT$_{\text{(NAACL'21)}}$ [77]	Faster R-CNN, ResNet-101	**74.2**	**92.4**	**96.0**	57.4	82.7	**89.9**	**492.6**
(Dense Local and Global Matching, without Pre-training)								
CAMP$_{\text{(ICCV'19)}}$ [5]	Faster R-CNN, ResNet-101	50.1	82.1	89.7	39.0	68.9	80.2	409.8
SCAN$_{\text{(ECCV'18)}}$ [3]	Faster R-CNN, ResNet-101	50.4	82.2	90.0	38.6	69.3	80.4	410.4
KASCE$_{\text{(IJCAI'19)}}$ [80]$*$	Faster R-CNN, ResNet-101	**56.6**	**84.5**	**92.0**	39.2	68.0	81.3	**421.6**
MPL$_{\text{(CVPR'20)}}$ [79]	Faster R-CNN, ResNet-101	46.9	77.7	87.6	34.4	64.2	75.9	386.7
GOT$_{\text{(ICML'20)}}$ [81]	Faster R-CNN, ResNet-101	50.5	80.2	89.8	38.1	66.8	78.5	403.9
IMRAM$_{\text{(CVPR'20)}}$ [18]	Faster R-CNN, ResNet-101	53.7	83.2	91.0	39.7	69.1	79.8	416.5
CAAN$_{\text{(CVPR'20)}}$ [6]	Faster R-CNN, ResNet-101	52.5	83.3	90.9	**41.2**	**70.3**	**82.9**	421.1
(Only Global Matching, without Pre-training)								
GVSE$_{\text{(AAAI'19)}}$ [82]	Faster R-CNN, ResNet-101	49.9	77.4	87.6	38.4	68.5	79.7	401.4
VSRN (ours)	Faster R-CNN, ResNet-101	53.0	81.1	89.4	40.5	70.6	81.1	415.7
VSRN++ (ours)	Faster R-CNN, ResNet-101	**54.7**	**82.9**	**90.9**	**42.0**	**72.2**	**82.7**	**425.4**

$*$ *indicates using human annotated scene graphs.*

TABLE 4
Comparisons of the Inference Time With Recent State-of-the-art Methods Whose Code is Publicly Available

Methods	Flickr30K 1K test set			MS-COCO 5-fold 1K test set			MS-COCO 5K test set		
	Encoding	Matching	Total	Encoding	Matching	Total	Encoding	Matching	Total
(Dense Local and Global Matching)									
SCAN$_{(ECCV'18)}$ [3]	9.7 s	599.0 s	608.7 s	44.6 s	2746.4 s	2791.0 s	44.2 s	14355.3 s	14399.5 s
BFAN$_{(MM'19)}$ [4]	12.9 s	1158.4 s	1171.3 s	58.7 s	5744.2 s	5802.9 s	58.6 s	32106.0 s	32164.6 s
CAMP$_{(ICCV'19)}$ [5]	4.3 s	1291.5 s	1295.8 s	19.9 s	6523.9 s	6543.8 s	20.7 s	40580.5 s	40601.2 s
MPL$_{(CVPR'20)}$ [79]	10.2 s	648.7 s	658.9 s	46.3 s	3021.0 s	3067.3 s	46.5 s	16661.8 s	16708.3 s
IMRAM$_{(CVPR'20)}$ [18]	9.8 s	680.5 s	690.3 s	47.7 s	3417.4 s	3465.1 s	48.5 s	17388.2 s	17436.7 s
(Only Global Matching)									
VSRN	16.7 s	4.7 s	21.4 s	74.3 s	21.6 s	95.9 s	74.4 s	115.4 s	189.8 s
VSRN++	20.4 s	4.9 s	25.3 s	97.3 s	20.9 s	118.2 s	103.6 s	116.2 s	219.8 s

"s" here represents for second. We list the time cost for calculating the embedding ("Encoding"), obtaining the similarity score for all image-text pairs ("Matching") and the total inference procedure ("Total"). Data loading time and feature extraction time are not included as they are the same for these methods. All methods are tested on the same machine using one single GPU and there is no code optimizing effort for fair comparisons (code of other methods is directly obtained from their official GitHub. Their code and ours are mainly build upon the official code base of VSE++ [7]). Our methods have demonstrated strong strength for the matching efficiency which is the key for the retrieval system especially when the encoding process can be done offline in advance. We also show VSRN++ does not bring heavy computation burden than VSRN while achieving much stronger retrieval performance.

helps to boost the retrieval performance, it is very time consuming at the model inference stage. In contrast, the proposed VSRN and VSRN++ perform reasoning on image region features and words to generate global representations for each image or text caption. At the matching stage, we follow the simple global matching strategy and only rely on the most basic inner product between the whole image representation and whole caption representation (only two vectors) to measure their similarity. Therefore, our methods are much more efficient than many recent works at the inference stage.

In Table 4, we report the running time comparisons with recent state-of-the-art methods whose code is publicly available. We conduct experiments on Flickr30K test set, MS-COCO 5-fold 1K test set and MS-COCO 5K test set, which are conducted on the same machine equipped with Intel Core^{TM} i7 CPU and NVIDIA TITAN XP GPUs. Regarding the total inference time, our methods (VSRN and VSRN++) are more than 30 times faster than other recent methods on Flickr30K and MS-COCO 5-fold 1K test sets. When testing on MS-COCO 5K test set, which is much larger than previous two sets our methods are more than 70 times faster than others. The larger the test set is, the larger the contrast of running time is. Although the inference time costs of other recent state-of-the-arts are not listed such as MPL [79], GOT [81], CAAN [6] because they have not released the source code, they actually build upon SCAN and should be slower than SCAN. Our methods also have strong strength for the matching efficiency which is the key for the retrieval system especially when the encoding process can be done offline in advance. Instead of following the recent trend of complex densely matching algorithms to obtain good performance while sacrificing the efficiency, we show that the simple global matching strategy can still be very effective, efficient and achieve even better performance based on our visual reasoning framework.

4.4 Ablation Studies

Analyze Each Component in the Model. We incrementally validate the effectiveness of each component in our model by conducting experiments on both Flickr30K test set and MS-COCO 5-fold 1K test set. As shown in the top six rows of Table 5, we first do ablation studies for the visual encoding part of our VSRN model. The text encoding part is kept the same as VSRN in these experiments, which includes one embedding layer described by Eq. (8) and one GRU layer. We start with a baseline model that does not perform any reasoning (noted as 'Visual FC + Mean-pool'). It simply adopts a mean-pooling layer on image region features $V = \{v_1, \ldots, v_k\}, v_i \in \mathbb{R}^D$, which is obtained after the visual embedding layers (Eq. (1)). Then we add one region relationship reasoning (RRR) layer (described in Section 3.4) before the mean-pooling operation into this baseline model and mark it as "Visual RRR". We also use the global semantic reasoning (GSR) module (described in Section 3.4) to replace the mean-pooling operation, which is noted as "Visual GSR". From results in Table 5, we validate that these two modules are both effective and help to obtain image representations that can be better matched with text captions. This is because RRR module can enhance feature representation with relationship information. These relationship enhanced features can further allow GSR module to perform global reasoning on a graph topology, which considers both current local region and global semantic correlations. We gradually add RRR layers before GSR and find improvements become less when adding more than 4 RRR layers. We also report results of VSRN trained without text generation loss L_G (noted as 4RRR+GSR*). Comparison shows that the joint optimization of matching and generation can help to improve around 2% relatively for R@1. In addition to the RRR, we further incorporate region location relationship reasoning (LRR) module described in Section 3.3 and validate its effectiveness. As shown in line 7 of the table, LRR brings around 2.7% relative improvement on Flickr dataset upon the already high performance. "4RRR +LRR+GSR" is taken as the final configuration of the visual encoding part for VSRN++.

For the textual encoding part, we follow a similar path and start with a fully connected layer followed with a mean-pooling operation to map the BERT feature to the same dimension as the visual feature. Although the BERT feature is known to have good ability of capturing long-

TABLE 5
Ablation Studies on the MS-COCO 5-Fold 1K Test set and Flickr 30K Test set

Methods	Flickr30K test set						MS-COCO 5-fold 1K test set					
	Caption Retrieval			Image Retrieval			Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Ablation for visual encoding with Faster R-CNN features												
Visual FC + Mean-pool	59.4	84.9	91.3	44.1	73.3	82.2	64.3	90.5	95.1	49.2	83.4	91.5
Visual RRR	67.3	89.0	94.0	48.8	78.0	86.2	68.5	93.2	96.3	56.8	87.2	94.2
Visual GSR	68.5	89.2	94.2	50.9	78.7	86.3	72.3	94.4	98.0	59.6	88.6	94.5
Visual 1RRR + GSR	70.3	89.5	94.2	52.4	79.1	86.4	75.3	94.7	98.1	62.1	89.2	94.9
Visual 4RRR + GSR	71.3	90.6	96.0	54.7	81.8	88.2	76.2	94.8	98.2	62.8	89.7	95.1
Visual 4RRR + GSR*	69.6	90.1	94.2	52.5	80.2	86.3	74.6	94.6	98.2	61.2	89.0	94.8
Visual 4RRR + LRR + GSR	73.2	93.0	96.2	56.6	83.5	89.4	76.8	95.2	98.3	63.3	90.0	95.2
Ablation for textual encoding with BERT features (Visual part using “4RRR + LRR + GSR”)												
Texual FC + Mean-pool	56.4	85.5	91.8	43.8	74.4	83.0	63.7	90.8	96.0	52.9	84.5	92.5
Texual WRR	63.2	87.1	93.3	47.5	76.8	85.2	66.2	92.1	96.7	55.5	86.1	93.1
Texual GRU	74.7	93.2	96.2	57.6	84.2	90.4	77.0	95.1	98.2	63.4	89.7	95.4
Texual WRR + GRU	76.8	94.0	96.9	58.7	85.0	90.8	77.5	95.5	98.3	63.8	89.9	95.8
Ablation for Sinkhorn-based Matching												
Without Sinkhorn	76.8	94.0	96.9	58.7	85.0	90.8	77.5	95.5	98.3	63.8	89.9	95.8
With Sinkhorn	79.2	94.6	97.5	60.6	85.6	91.4	77.9	96.0	98.5	64.1	91.0	96.1

“” means model training without using text generation loss L_G.*

term complex relations of words in the sentence with self-attention, we find it is not ideal if only relying on the fixed BERT feature with such simple mapping functions. We then gradually add word-relationship reasoning (WRR) and textual GRU modules. From bottom parts of Table 5, we find both of them are very effective to help improve the matching results. We have also tried more layers of WRR followed with one GRU and found their results are comparable with single WRR + GRU. Therefore, single layer WRR followed with one GRU is taken as the final configuration of the textual encoding part for VSRN++. In the last two rows, we have also studied the effectiveness of Sinkhorn-based matching algorithm described in Section 3.6.

Effects of Region Ordering. Since our GSR module (Section 3.4) sequentially processes region features and generates the representation of the whole image gradually, we consider several choices about region ordering for this process in Table 6 and report mean of 3 runs. One possible setting (VSRN-Confidence) is the descending order of their class detection confidence scores that are generated by the bottom-up attention detector. We expect this to encourage the model to focus on the easy regions with high confidence first and then inferring more difficult regions based on the semantic context. We also try the way of feeding regions in the reverse order of confidence, which is marked as “VSRN-Confidence-R”. Another option (VSRN-BboxSize) is to sort the detection bounding boxes of these regions in descending order, as this lets the model to obtain global scene information first. We also test the model with randomly ordering of the regions (VSRN-Random). Results in Table 6 show that reasoning in a specific order can help improve the performance than the random one. VSRN-Confidence, VSRN-Confidence-R and VSRN-BboxSize achieve comparable results with a reasonable ordering scheme. We take VSRN-Confidence as the setting of VSRN in our previous experiments. Besides, we also find the variance of R@1 is around 1 point for these different settings, which suggests VSRN is robust to the ordering scheme used. One possible reason could be that global information is constructed during the region relationship reasoning step. Based on these relationship enhanced feature, GSR can be then performed on global graph topologies.

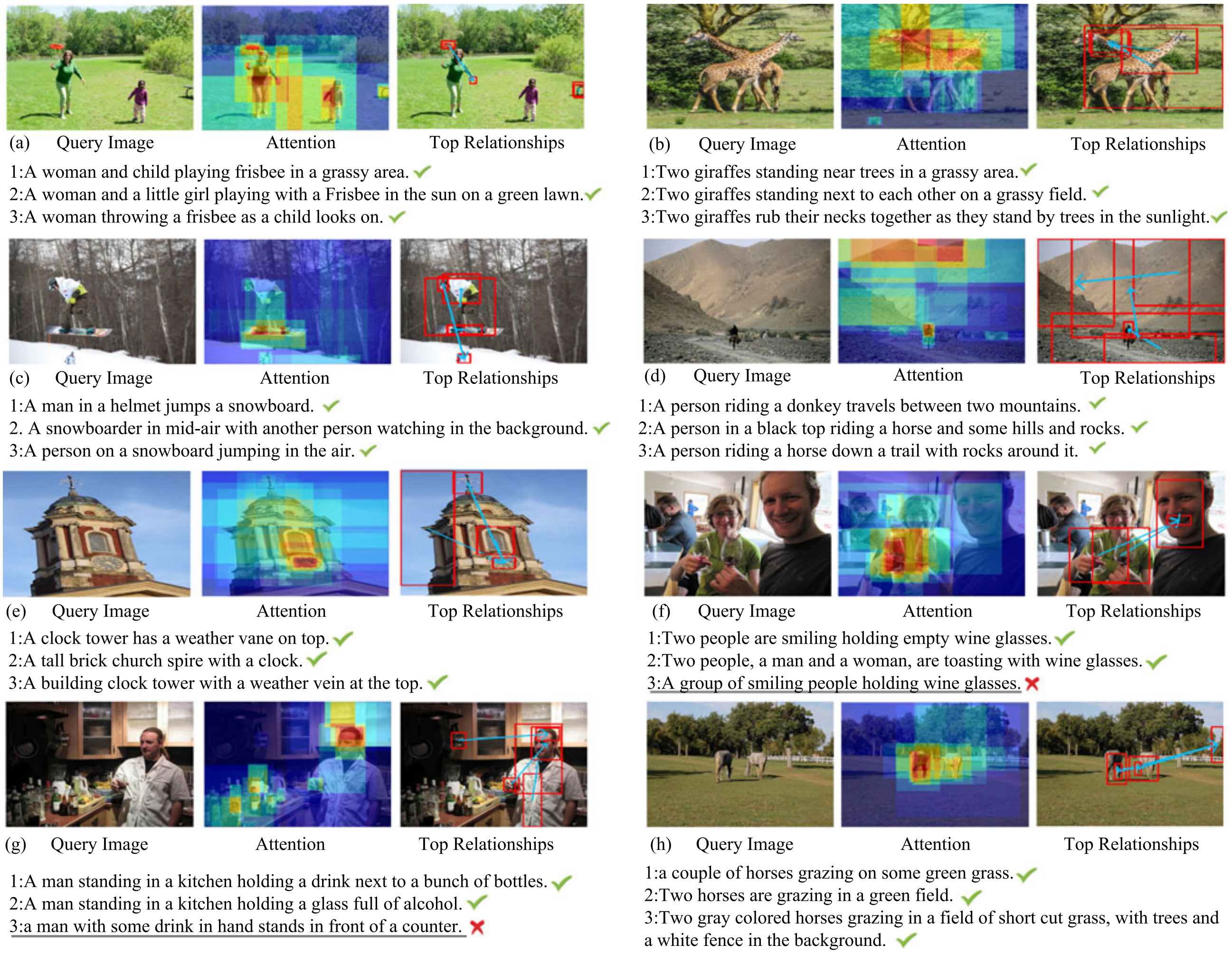
TABLE 6
Studies to Analyze the Effects of Region Ordering

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
VSRN-Random	74.8	94.0	97.8	61.6	89.0	94.5
VSRN-BboxSize	75.6	94.5	98.2	62.4	89.4	94.8
VSRN-Confidence-R	75.8	94.9	98.3	62.6	89.7	95.1
VSRN-Confidence	76.0	94.7	98.2	62.7	89.6	95.1

Results are reported in terms of mean Recall@K (R@K) over 3 runs. “VSRN-Confidence-R” represents in reverse order of confidence.

4.5 Visualization and Analysis

Attention Visualization of the Final Image Representation. Learning to generate the image representation that captures key semantic concepts and objects in a given image scene is one important goal of the proposed visual semantic reasoning model. To verify this, we design an attention-based visualization method to show the correlations between the final whole image representation and these region representations within this image. Specifically, we calculate the inner product similarity (same as in Eq. (12)) between each region feature $V^* = \{v_1^*, \dots, v_k^*\}, v_i^* \in \mathbb{R}^D$ and the final whole image representation $I \in \mathbb{R}^D$. Then we rank the image regions V^* in the descending order of their correlation with I and assign a ranking score s_i to each v_i^* according to its rank r_i. The score is calculate by $s_i = \lambda(k - r_i)^2$, where k is the total number of regions, r_i is the corresponding rank, λ is a parameter used to emphasize the high ranked regions and is set 50 in our experiments. To generate the final attention map (similarity map), the attention score at each pixel



(a) Query Image Attention Top Relationships

1:A woman and child playing frisbee in a grassy area. ✓
2:A woman and a little girl playing with a Frisbee in the sun on a green lawn.✓
3:A woman throwing a frisbee as a child looks on. ✓

(b) Query Image Attention Top Relationships

1:Two giraffes standing near trees in a grassy area.✓
2:Two giraffes standing next to each other on a grassy field. ✓
3:Two giraffes rub their necks together as they stand by trees in the sunlight.✓

(c) Query Image Attention Top Relationships

1:A man in a helmet jumps a snowboard. ✓
2. A snowboarder in mid-air with another person watching in the background. ✓
3:A person on a snowboard jumping in the air. ✓

(d) Query Image Attention Top Relationships

1:A person riding a donkey travels between two mountains. ✓
2:A person in a black top riding a horse and some hills and rocks. ✓
3:A person riding a horse down a trail with rocks around it. ✓

(e) Query Image Attention Top Relationships

1:A clock tower has a weather vane on top. ✓
2:A tall brick church spire with a clock. ✓
3:A building clock tower with a weather vein at the top. ✓

(f) Query Image Attention Top Relationships

1:Two people are smiling holding empty wine glasses. ✓
2:Two people, a man and a woman, are toasting with wine glasses. ✓
3:A group of smiling people holding wine glasses. ✗

(g) Query Image Attention Top Relationships

1:A man standing in a kitchen holding a drink next to a bunch of bottles. ✓
2:A man standing in a kitchen holding a glass full of alcohol. ✓
3:a man with some drink in hand stands in front of a counter. ✗

(h) Query Image Attention Top Relationships

1:a couple of horses grazing on some green grass. ✓
2:Two horses are grazing in a green field. ✓
3:Two gray colored horses grazing in a field of short cut grass, with trees and a white fence in the background. ✓

Fig. 3. Qualitative results of the image-to-text retrieval for VSRN on MS-COCO test set. For each image query, we show the top-3 ranked text caption. Ground-truth matched sentences are with check marks, while some sentences sharing similar meanings as ground-truth ones are marked with gray underline. We also show the attention visualization of the final image representation besides its corresponding image. Our model generates interpretable image representation that captures key objects and semantic concepts in the scene. (Best viewed in color when zoomed in.)

location is calculated by adding up ranking scores of these regions this pixel belongs to. With this visualization way, we can show visual attention maps when obtain qualitative results of text-to-image (image) retrieval and image-to-text (caption) retrieval.

Qualitative Results. In Fig. 3, for given image queries on MS-COCO test set, we show the top-3 retrieved captions for each image query obtained by VSRN. The retrieved captions are listed in order of similarity scores predicted by VSRN. From these results, we find that even for image queries with complex and cluttered scenes, our model can retrieve correct results in the top ranked captions. We also find the model outputs some reasonable mismatches, such as (f)-3 and (g)-3. Although these retrieved captions are not labeled as the ground-truth pair with the given image queries, they also describe some key parts of the scene. In addition to the retrieval results, we visualize attention maps obtained by the method described in the last paragraph. From the attention visualization, we can find that VSRN generates image representations that capture key semantic concepts and objects in the scene.

Besides, we are also interested to find out whether the model learns some reasonable semantic relationships during the reasoning process. Therefore, for each query image, we further show the top three edges in the first RRR layer as well as image region nodes connected by these edges. These visualizations are noted as “Top Relationships”, where edges are represented by blue lines with arrows and image regions are noted by red bounding boxes. We can find that the model can capture semantic relationships such as “human hand” and “frisbee”, “wine glassesd” and “human mouth”, “horses” and “grass/trees”. Other reasonable correlations such as relationships between different body parts of humans or animals. The behaviors of the model are more interpretable based on the reasoning mechanism. Without using strong supervision such as scene graph or knowledge graph, the model can capture some reasonable semantic relationships when learning the alignments between images and captions.

Qualitative Comparisons Between VSRN and VSRN++. In Fig. 4, we further show qualitative comparisons between VSRN and VSRN++ on MS-COCO test set for image-to-text retrieval (top) and text-to-image retrieval (bottom). For the challenge cases that VSRN fails, VSRN++ can well retrieve the ground truth images or text captions in the top-3 list. This is mainly due to the better image-text embedding space learnt by VSRN++. From the attention visualizations, we also find image representations generated by VSRN++ can

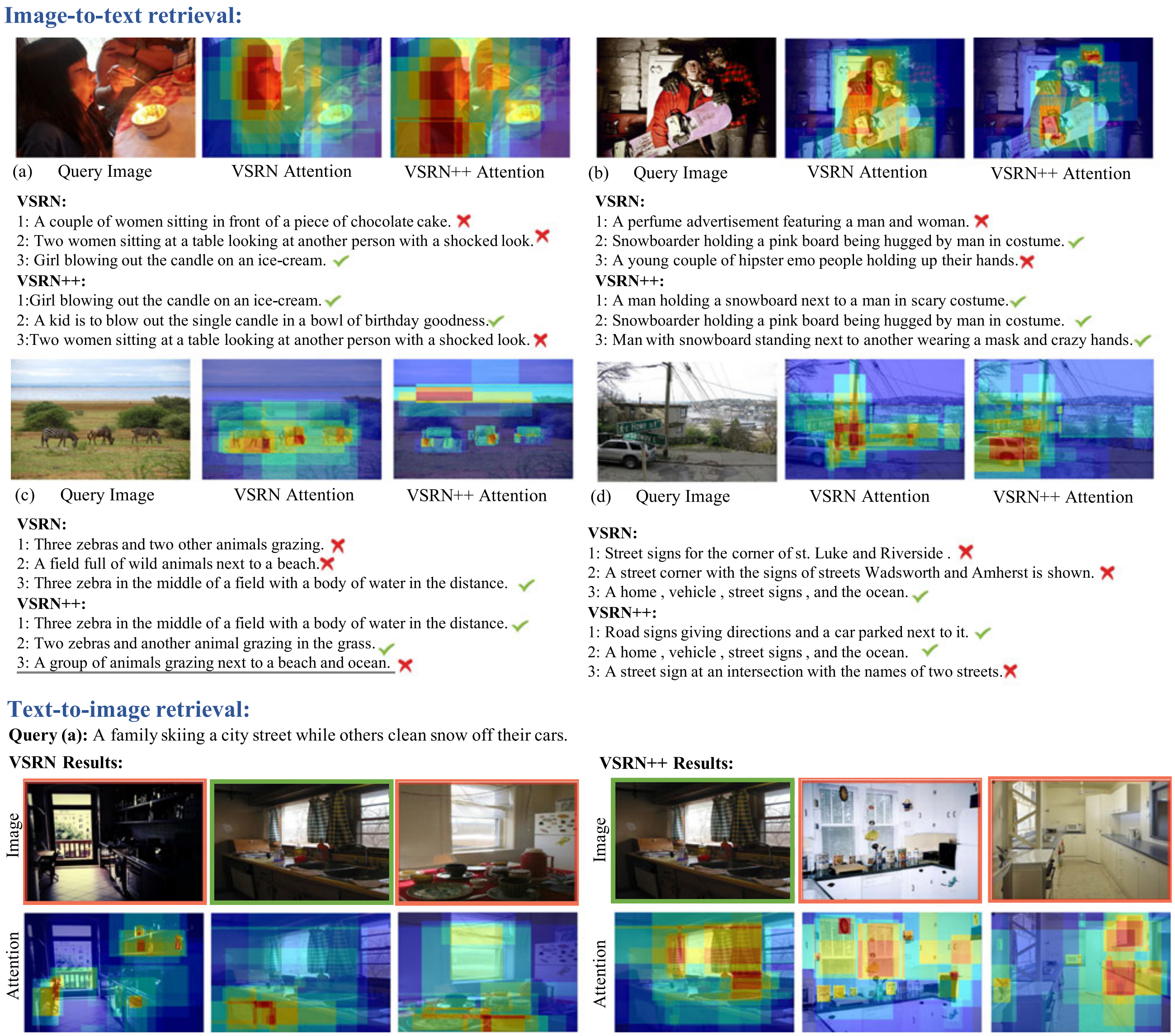


Image-to-text retrieval:

(a) Query Image VSRN Attention VSRN++ Attention

VSRN:
1: A couple of women sitting in front of a piece of chocolate cake. ✗
2: Two women sitting at a table looking at another person with a shocked look. ✗
3: Girl blowing out the candle on an ice-cream. ✓
VSRN++:
1:Girl blowing out the candle on an ice-cream. ✓
2: A kid is to blow out the single candle in a bowl of birthday goodness. ✓
3:Two women sitting at a table looking at another person with a shocked look. ✗

(b) Query Image VSRN Attention VSRN++ Attention

VSRN:
1: A perfume advertisement featuring a man and woman. ✗
2: Snowboarder holding a pink board being hugged by man in costume. ✓
3: A young couple of hipster emo people holding up their hands. ✗
VSRN++:
1: A man holding a snowboard next to a man in scary costume. ✓
2: Snowboarder holding a pink board being hugged by man in costume. ✓
3: Man with snowboard standing next to another wearing a mask and crazy hands. ✓

(c) Query Image VSRN Attention VSRN++ Attention

VSRN:
1: Three zebras and two other animals grazing. ✗
2: A field full of wild animals next to a beach. ✗
3: Three zebra in the middle of a field with a body of water in the distance. ✓
VSRN++:
1: Three zebra in the middle of a field with a body of water in the distance. ✓
2: Two zebras and another animal grazing in the grass. ✓
3: A group of animals grazing next to a beach and ocean. ✗

(d) Query Image VSRN Attention VSRN++ Attention

VSRN:
1: Street signs for the corner of st. Luke and Riverside . ✗
2: A street corner with the signs of streets Wadsworth and Amherst is shown. ✗
3: A home , vehicle , street signs , and the ocean. ✓
VSRN++:
1: Road signs giving directions and a car parked next to it. ✓
2: A home , vehicle , street signs , and the ocean. ✓
3: A street sign at an intersection with the names of two streets. ✗

Text-to-image retrieval:

Query (a): A family skiing a city street while others clean snow off their cars.

VSRN Results: **VSRN++ Results:**

Fig. 4. Qualitative comparisons between VSRN and VSRN++ for image-to-text retrieval and text-to-image retrieval on MS-COCO test set. For each image query, we show the top-3 ranked text caption, where ground-truth matched sentences are with check marks. For each text query, we show the top-3 retrieved images, ranking from left to right. The true matches are outlined in green boxes and attention visualizations are shown.

not only focus on the key objects but also cover important contexts that help to describe the image in a more fine-grained way. For example the image-to-text retrieval, VSRN ++ pay more attentions on the candles in (a), the mask of the man behind in (b), the water area in (c) and the car in (d), which helps VSRN++ better handle these challenge cases than VSRN. We have similar observations for the text-to-image retrieval results.

5 Conclusion

In this paper, we present an intuitive and interpretable model to learn a common embedding space for alignments between images and text descriptions. To address the visual semantic discrepancy, we perform region relationship reasoning and global semantic reasoning to generate overall visual and textual representations. The enhanced image representation can capture key objects as well as semantic concepts in a scene, so that it can be better matched with the corresponding text descriptions. We have conducted extensive experiments on Flickr30K dataset and MS-COCO dataset to evaluate the effectiveness of the proposed model. Results show that our method outperforms several recent state-of-the-arts for the caption retrieval and image retrieval tasks. We also incrementally validate effectiveness of each component in our model by conducting ablation studies on two datasets.

In addition to the effectiveness, our methods are also very efficient at the inference stage. We argue this efficiency is crucial when evaluating the model performance especially when the model has potential to be used in search engines for a large-scale database towards image or text queries. Benefited from the effective overall representation learning with visual semantic reasoning, our method can already achieve very strong performance only relying on the classic inner-product to obtain similarity scores between images and texts. This is much more efficient than the recent popular methods focusing on complex matching strategies. Experiments on MS-

COCO 5K test set show that VSRN and VSRN++ are more than 70 times faster than recent state-of-the-art methods with code available. Instead of following the recent trend to pursue good performance while sacrificing the efficiency, we show that the classic global matching strategy can still be very effective, efficient and achieve even better performance based on our visual semantic reasoning framework.

We further investigate the interpretability of our model. The proposed attention-based visualization strategy elegantly addresses a common problem for image-level similarity measure, which traces back to the individual patches about the significance. From the attention visualization, we validate that our model can generate image representations that capture key semantic concepts and objects in the scene. Besides, we further visualize the top edges in the relationship reasoning layer and have interesting findings from the visualizations. Without using strong supervision such as scene graph or knowledge graph, the model can capture some reasonable semantic relationships when learning the alignments between images and captions.

In the future it may be illuminating to deploy our method on other tasks related to vision-language domains and learning interpretable representations by reasoning. Based on our explorations, we also expect more future work can further study the reasoning model for representation learning with clear rules and sophisticated logic.

References

[1] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014.

[3] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 212–228.

[4] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, and Y. Zhang, "Focus your attention: A bidirectional focal attention network for image-text matching," in *Proc. 27th ACM Int. Conf.*, 2019, pp. 3–11.

[5] Z. Wang *et al.*, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5763–5772.

[6] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3533–3542.

[7] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–14.

[8] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1969–1978.

[9] K. Li, C. Fang, Z. Wang, S. Kim, H. Jin, and Y. Fu, "Screencast tutorial video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12523–12532.

[10] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6964–6974.

[11] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1855–1865.

[12] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.

[13] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[14] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *Neuroscientist*, 2014, pp. 509–521.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv*.

[18] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12652–12660.

[19] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 4653–4661.

[20] M. Fey, J. E. Lenssen, C. Morris, J. Masci, and N. M. Kriege, "Deep graph matching consensus," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–16.

[21] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4937–4946.

[22] G. Peyré *et al.*, "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, pp. 355–607, 2019.

[23] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[24] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Trans. Assoc. Comput. Linguistics*, vol. 32, pp. 595–603, 2015.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[26] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–12.

[27] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3441–3450.

[28] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1473–1482.

[29] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–9.

[30] J. Wehrmann, C. Kolling, and R. C. Barros, "Adaptive cross-modal embeddings for image-text alignment," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 12 313–12 320.

[31] Y. Huang and L. Wang, "ACMM: Aligned cross-modal memory for few-shot image and sentence matching," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5773–5782.

[32] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[33] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2020, pp. 11 336–11 344.

[34] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10434–10413.

[35] Y.-C. Chen *et al.*, "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[36] X. Li *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[38] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9215–9523.

[39] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3128–3137.

[40] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6163–6171.

[41] C. Cao *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2956–2964.

[42] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Representations Workshop*, 2014, pp. 1–8.

[43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[44] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 1084–1102.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[46] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Guided attention inference network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2996–3010, Dec. 2020.

[47] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Attention bridging network for knowledge transfer," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5197–5206.

[48] A. Newell, "Physical symbol systems," *Cogn. Sci.*, vol. 4, pp. 135–183, 1980.

[49] J. R. Hobbs, M. E. Stickel, and P. Martin, "Interpretation as abduction," *Artif. Intell.*, vol. 63, pp. 69–142, 1993.

[50] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, pp. 335–346, 1990.

[51] N. Lao, T. Mitchell, and W. W. Cohen, "Random walk inference and learning in a large scale knowledge base," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2011, pp. 529–539.

[52] Y. Zhang *et al.*, "Multimodal style transfer via graph cuts," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5943–5951.

[53] S. Chandra, N. Usunier, and I. Kokkinos, "Dense and low-rank gaussian CRFs using deep embeddings," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5113–5122.

[54] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6137–6145.

[55] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 684–699.

[56] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.

[57] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–14.

[58] W. Yu, J. Zhou, W. Yu, X. Liang, and N. Xiao, "Heterogeneous graph learning for visual commonsense reasoning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 2769–2779.

[59] A. Wu, L. Zhu, Y. Han, and Y. Yang, "Connective cognition network for directional visual commonsense reasoning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5669–5679.

[60] W. Norcliffe-Brown , S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8344–8353.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[62] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.

[63] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7239–7248.

[64] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[65] I. Kokkinos, M. M. Bronstein, R. Litman, and A. M. Bronstein, "Intrinsic shape context descriptors for deformable shapes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 159–166.

[66] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5425–5434.

[67] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 60–65.

[68] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[69] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10938–10947.

[70] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific J. Math.*, vol. 21, pp. 343–348, 1967.

[71] R. P. Adams and R. S. Zemel, "Ranking via sinkhorn propagation," 2011, *arXiv:1106.1925*.

[72] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.

[73] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[74] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4534–4542.

[75] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4127–4136.

[76] Y. Huang, Q. Wu, W. Wang, and L. Wang, "Image and sentence matching via semantic concepts and order learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 636–650, Mar. 2020.

[77] S. Sun, Y.-C. Chen, L. Li, S. Wang, Y. Fang, and J. Liu, "LightningDOT: Pre-training visual-semantic embeddings for real-time image-text retrieval," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2021, pp. 982–997.

[78] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6616–6628.

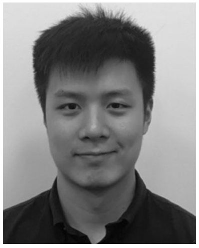
[79] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13002–13011.

[80] B. Shi, L. Ji, P. Lu, Z. Niu, and N. Duan, "Knowledge aware semantic concept expansion for image-text matching," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 5182–5189.

[81] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu, "Graph optimal transport for cross-domain alignment," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1542–1553.

[82] Y. Huang, Y. Long, and L. Wang, "Few-shot image and sentence matching via gated visual-semantic embedding," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2019, pp. 8489–8496.

[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv*.

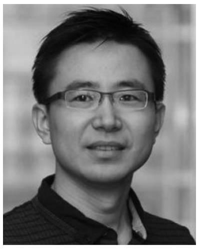


Kunpeng Li (Member, IEEE) received the BEng degree in information engineering from the South China University of Technology, China, and the PhD degree in computer engineering from Northeastern University, Boston, MA. He is currently a research scientist with Facebook Reality Labs, CA. He has also spent time with Google Research and Adobe Research, as a research intern. His research interests include learning with limited supervision, scene understanding, vision & language, and video understanding.

Yulun Zhang (Member, IEEE) received the BE degree from the School of Electronic Engineering, Xidian University, China, in 2013, the ME degree from the Department of Automation, Tsinghua University, China, in 2017, and the PhD degree from the Department of ECE, Northeastern University, USA, in 2021. He is currently a postdoctoral researcher with Computer Vision Lab, ETH Zürich, Switzerland. He was also a research fellow with Harvard University. His research interests include image or video restoration and synthesis, biomedical image analysis, model compression, and computational imaging. He was the recipient of the Best Student Paper Award at VCIP in 2015 and the Best Paper Award at ICCV RLQ Workshop in 2019.

Kai Li (Member, IEEE) received the BEng and MEng degrees from Wuhan University, Wuhan, China, in 2014 and 2016, respectively, and the PhD degree from Northeastern University, Boston, MA, USA. Since 2021, he has been a researcher with Machine Learning Department, NEC Laboratories America Inc., Princeton, NJ, USA. His research interests include machine learning and computer vision, particularly in domain adaptation, and zero/few-shot learning.

Yuanyuan Li (Student Member, IEEE) received the BE degree from the School of Electronic and Information Engineering, South China University of Technology, China, and the MS degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University. She is currently working toward the PhD degree in computer engineering with Northeastern University, Boston, USA. Her research interests include networking, optimization, and machine learning. She is an ACM student member.



Yun Fu (Fellow, IEEE) received the BEng degree in information engineering and the MEng degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, and the MS degree in statistics and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign. Since 2012, he has been an interdisciplinary faculty member affiliated with the College of Engineering and the College of Computer and Information Science, Northeastern University. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. His research interests include machine learning, computational intelligence, Big Data mining, computer vision, pattern recognition, and cyber-physical systems. He is an associate editor, chairs, PC member and reviewer of many top journals and international conferences/workshops. He was the recipient of the seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation, 12 best paper awards from IEEE, IAPR, SPIE, SIAM, many main Industrial Research awards from Google, Amazon, Samsung, JPMorgan Chase, Cisco and Adobe. He is currently an associate editor for *IEEE Transactions on Image Processing* (TIP). He is a fellow of IAPR, OSA, and SPIE, a Lifetime distinguished member of ACM, Lifetime senior member of the AAAI and Institute of Mathematical Statistics, a member of the ACM Future of Computing Academy, Global Young Academy, AAAS, INNS and Beckman Graduate Fellow during 2007–2008.