

# Visual Semantic Reasoning for Image-Text Matching

Kunpeng Li<sup>1</sup>, Yulun Zhang<sup>1</sup>, Kai Li<sup>1</sup>, Yuanyuan Li<sup>1</sup> and Yun Fu<sup>1,2</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, Northeastern University, Boston, MA

<sup>2</sup>Khoury College of Computer Science, Northeastern University, Boston, MA

## Abstract

*Image-text matching has been a hot research topic bridging the vision and language areas. It remains challenging because the current representation of image usually lacks global semantic concepts as in its corresponding text caption. To address this issue, we propose a simple and interpretable reasoning model to generate visual representation that captures key objects and semantic concepts of a scene. Specifically, we first build up connections between image regions and perform reasoning with Graph Convolutional Networks to generate features with semantic relationships. Then, we propose to use the gate and memory mechanism to perform global semantic reasoning on these relationship-enhanced features, select the discriminative information and gradually generate the representation for the whole scene. Experiments validate that our method achieves a new state-of-the-art for the image-text matching on MS-COCO [28] and Flickr30K [39] datasets. It outperforms the current best method by 6.8% relatively for image retrieval and 4.8% relatively for caption retrieval on MS-COCO (Recall@1 using 1K test set). On Flickr30K, our model improves image retrieval by 12.6% relatively and caption retrieval by 5.8% relatively (Recall@1). Our code is available at <https://github.com/KunpengLi1994/VSRN>.*

## 1. Introduction

Vision and language are two important aspects of human intelligence to understand the real world. A large amount of research [5, 9, 23] has been done to bridge these two modalities. Image-text matching is one of the fundamental topics in this field, which refers to measuring the visual-semantic similarity between a sentence and an image. It has been widely adopted to various applications such as the retrieval of text descriptions from image queries or image search for given sentences.

Although a lot of progress has been achieved in this area, it is still a challenge problem due to the huge visual semantic discrepancy. When people describe what they see in the

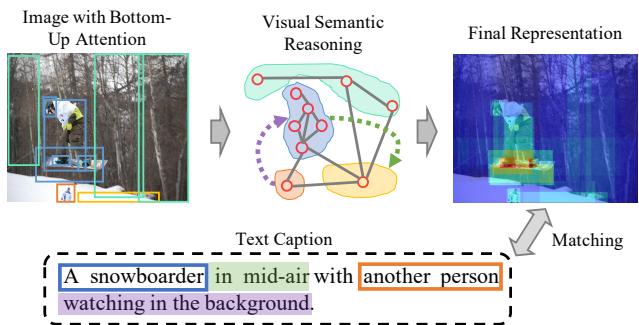


Figure 1. The proposed Visual Semantic Reasoning Network (VSRN) performs reasoning on the image regions to generate representation for an image. The representation captures key objects (boxes in the caption) and semantic concepts (highlight parts in the caption) of a scene as in the corresponding text caption.

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 Figure 1). Visual reasoning about objects and semantics is crucial for humans during this process. However, the current existing visual-text matching systems lack such kind of reasoning mechanism. Most of them [5] represent concepts in an image by Convolutional Neural Network (CNN) features extracted by convolutions with a specific receptive field, which only perform local pixel-level analysis. It is hard for them to recognize the high-level semantic concepts. More recently, [23] make use of region-level features from object detectors and discover alignments between image regions and words. Although grasping some local semantic concepts within regions including multiple objects, these methods still lack the global reasoning mechanism that allows information communication between regions farther away.

To address this issue, we propose Visual Semantic Reasoning Network (VSRN) to generate visual representation that captures both objects and their semantic relationships. We start from identifying salient regions in images by following [1, 23]. In this way, salient region detection at stuff/object level can be analogized to the bottom-up at-

tention that is consistent with human vision system [16]. Practically, the bottom-up attention module is implemented using Faster R-CNN [33].<sup>2</sup> We then build up connections between these salient regions and perform reasoning with Graph Convolutional Networks (GCN) [18] to generate features with semantic relationships.

Different image regions and semantic relationships would have different contributions for inferring the image-text similarity and some of them are even redundant. Therefore, we further take a step to attend important ones when generating the final representation for the whole image. We propose to use the gate and memory mechanism [3] 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 final image representation captures more key semantic concepts than those from existing methods that lack a reasoning mechanism, therefore, can help to achieve better image-text matching performance.

In addition to quantitative evaluation of our model on standard benchmarks, we also design an interpretation method to analyze what has been learned inside the reasoning model. Correlations between the final image representation and each region feature are visualized in an attention format. As shown in Figure 1, we find the learned image representation has high response at these regions that include key semantic concepts.

To sum up, our main contributions are: (a) We propose a simple and interpretable reasoning model VSRN to generate enhanced visual representations by region relationship reasoning and global semantic reasoning. (b) We design an interpretation method to visualize and validate that the generated image representation can capture key objects and semantic concepts of a scene, so that it can be better aligned with the corresponding text caption. (c) The proposed VSRN achieves a new state-of-the-art for the image-text matching on MS-COCO [28] and Flickr30K [39] datasets. Our VSRN outperforms the current best method SCAN [23] by 6.8% relatively for image retrieval and 4.8% relatively for caption retrieval on MS-COCO (Recall@1 using 1K test set). On Flickr30K, our model improves image retrieval by 12.6% relatively and caption retrieval by 5.8% relatively (Recall@1).

## 2. Related Work

**Image-Text Matching.** Our work is related to existing methods proposed for image-text matching, where the key issue is measuring the visual-semantic similarity between a text and an image. Learning a common space where text and image feature vectors are comparable is a typical solution for this task. Frome et al. [6] propose a feature em-

bedding framework that uses Skip-Gram and CNN to extract feature representations for cross-modal. Then a ranking loss is adopted to encourage the distance between the mismatched image-text pair is larger than that between the matched pair. Kiros et al. [19] use a similar framework and adopt LSTM [12] instead of Skip-Gram for the learning of text representations. Vendrov et al. [35] design a new objective function that encourages the order structure of visual semantic can be preserved hierarchy. Faghri et al. [5] focus more on hard negatives and obtain good improvement using a triplet loss. Gu et al. [8] further improve the learning of cross-view feature embedding by incorporating generative objectives. Our work also belongs to this direction of learning joint space for image and sentence with an emphasis on improving image representations.

**Attention Mechanism.** Our work is also inspired by bottom-up attention mechanism and recent image-text matching methods based on it. Bottom-up attention [16] refers to salient region detection at stuff/object level can be analogized to the spontaneous bottom-up attention that is consistent with human vision system [16, 24–27]. Similar observation has motivated other existing work. In [15], R-CNN [7] is adopted to detect and encode image regions at object level. Image-text similarity is then obtained by aggregating all word-region pairs similarity scores. Huang et al. [14] 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. [23] further propose an attention model towards attending key words and image regions for predicting the text-image similarity. Following them, we also start from bottom-up region features of an image. However, to the best of our knowledge, no study has attempted to incorporate global spatial or semantic reasoning when learning visual representations for image-text matching.

**Relational Reasoning Methods.** Symbolic approaches [31] are the earliest form of reasoning in artificial intelligence. In these methods, relations between symbols are represented by the form of logic and mathematics, reasoning happens by abduction and deduction [11] etc. However, in order to make these systems can be used practically, symbols need to be grounded in advance. More recent methods, such as path ranking algorithm [22], perform reasoning on structured knowledge bases by taking use of statistical learning to extract effective patterns. As an active research area, graph-based methods [40] have been very popular in recent years and shown to be an efficient way of relation reasoning. Graph Convolution Networks (GCN) [18] are proposed for semi-supervised classification. Yao et al. [38] train a visual relationship detection model on Visual Genome dataset [21] and use a GCN-based encoder to encode the detected relationship information into an image

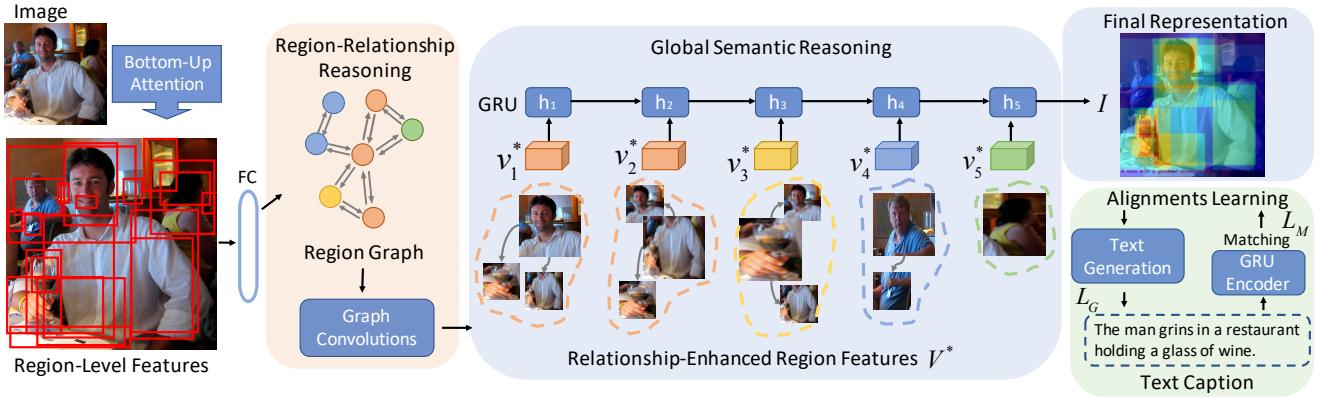


Figure 2. An overview of the proposed Visual Semantic Reasoning Network (VSRN). Based on salient image regions from bottom-up attention (Sec. 3.1), VSRN first performs region relationship reasoning on these regions using GCN to generate features with semantic relationships (Sec. 3.2). Then VSRN takes use of the gate and memory mechanism to perform global semantic reasoning on the relationship enhanced features, select the discriminative information and gradually generate the representation for the whole scene (Sec. 3.3). The whole model is trained with joint optimization of matching and sentence generation (Sec. 3.4). The attention of the representation (top right) is obtained by calculating correlations between the final image representation and each region feature (Sec. 4.5).

captioning framework. Yang et al. [37] utilize GCNs to incorporate the prior knowledge into a deep reinforcement learning framework improve semantic navigation in unseen scenes and towards novel objects. We also adopt the reasoning power of graph convolutions to obtain image region features enhanced with semantic relationship. But we do not need extra database to build the relation graph (e.g. [38] needs to train the relationship detection model on Visual Genome). Beyond this, we further perform global semantic reasoning on these relationship-enhanced features, so that the final image representation can capture key objects and semantic concepts of a scene.

### 3. Learning Alignments with Visual Semantic Reasoning

We describe the detail structure of the Visual Semantic Reasoning Network (VSRN) for image-text matching in this section. 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 [1] (Sec. 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 (Sec. 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 (Sec. 3.3). For the text caption part, we learn a representation for the sentence using RNNs. Finally, the whole model is trained with joint optimization of image-sentence

matching and sentence generation (Sec. 3.4).

#### 3.1. Image Representation by Bottom-Up Attention

Taking the advantage of bottom-up attention [1], each image can be represented by a set of features  $V = \{v_1, \dots, v_k\}$ ,  $v_i \in \mathbb{R}^D$ , such that each feature  $v_i$  encodes an object or a salient region in this image. Following [1, 23], we implement the bottom-up attention with a Faster R-CNN [33] model using ResNet-101 [10] as the backbone. It is pre-trained on the Visual Genomes dataset [21] by [1]. 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 [1, 23]. 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:

$$v_i = W_f f_i + b_f. \quad (1)$$

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

### 3.2. Region Relationship Reasoning

Inspired by recent advances in deep learning based visual reasoning [2, 34, 41], we build up a region relationship reasoning model 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.

$$R(v_i, v_j) = \varphi(v_i)^T \phi(v_j), \quad (2)$$

where  $\varphi(v_i) = W_\varphi v_i$  and  $\phi(v_j) = W_\phi v_j$  are two embeddings. The weight parameters  $W_\varphi$  and  $W_\phi$  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) [18] 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:

$$V^* = W_r(RVW_g) + V, \quad (3)$$

where  $W_g$  is the weight matrix of the GCN layer with dimension of  $D \times D$ .  $W_r$  is the weight matrix of 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  $V^* = \{v_1^*, \dots, v_k^*\}$ ,  $v_i^* \in \mathbb{R}^D$  is the relationship enhanced representation for image region nodes.

### 3.3. 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 by putting the sequence of region features  $V^* = \{v_1^*, \dots, v_k^*\}$ ,  $v_i^* \in \mathbb{R}^D$ , one by one into GRUs [3]. The description of the whole scene will gradually grow and update in the memory cell (hidden state)  $m_i$  during this reasoning process.

At each reasoning 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:

$$z_i = \sigma_z(W_z v_i^* + U_z m_{i-1} + b_z), \quad (4)$$

where  $\sigma_z$  is a sigmoid activation function.  $W_z$ ,  $U_z$  and  $b_z$  are weights and bias.

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

$$\tilde{m}_i = \sigma_m(W_m v_i^* + U_z(r_i \circ m_{i-1}) + b_m), \quad (5)$$

where  $\sigma_m$  is a tanh activation function.  $W_m$ ,  $U_m$  and  $b_m$  are weights and bias.  $\circ$  is an element-wise multiplication.  $r_i$  is the reset gate that decides what content to forget based on the reasoning between  $v_i^*$  and  $m_{i-1}$ .  $r_i$  is computed similarly to the update gate as:

$$r_i = \sigma_r(W_r v_i^* + U_r m_{i-1} + b_r), \quad (6)$$

where  $\sigma_r$  is a sigmoid activation function.  $W_r$ ,  $U_r$  and  $b_r$  are weights and bias.

Then the description of the whole scene  $m_i$  at the current step is a linear interpolation using update gate  $z_i$  between the previous description  $m_{i-1}$  and the new content  $\tilde{m}_i$ :

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

where  $\circ$  is an element-wise multiplication. Since each  $v_i^*$  includes global relationship information, update of  $m_i$  is actually based on reasoning on a graph topology, which considers both current local region and global semantic correlations. We take the memory cell  $m_k$  at the end of the sequence  $V^*$  as the final representation  $I$  for the whole image, where  $k$  is the length of  $V^*$ .

### 3.4. Learning Alignments by Joint Matching and Generation

To connect vision and language domains, we use a GRU-based text encoder [3, 5] to map the text caption to the same  $D$ -dimensional semantic vector space  $C \in \mathbb{R}^D$  as the image representation  $I$ , which considers semantic context in the sentence. Then we jointly optimize matching and generation to learn the alignments between  $C$  and  $I$ .

For the matching part, we adopt a hinge-based triplet ranking loss [5, 15, 23] with emphasis on hard negatives [5], i.e., the negatives closest to each training query. We define the loss as:

$$L_M = [\alpha - S(I, C) + S(I, \hat{C})]_+ + [\alpha - S(I, C) + S(\hat{I}, C)]_+, \quad (8)$$

where  $\alpha$  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 function in the joint embedding space. We use the usual inner product as  $S(\cdot)$  in our experiments.  $\hat{I} = \arg \max_{j \neq I} S(j, C)$  and  $\hat{C} = \arg \max_{d \neq C} S(I, d)$  are the hardest negatives for a positive pair  $(I, C)$ . For computational efficiency, instead of finding the hardest negatives in the entire training set, we find them within each mini-batch.

For the generation part, the learned visual representation should also have the ability to generate sentences that are close to the ground-truth captions. Specifically, we use a sequence to sequence model with attention mechanism [36] to achieve this. We maximize the log-likelihood of the predicted output sentence. The loss function is defined as:

$$L_G = - \sum_{t=1}^l \log p(y_t | y_{t-1}, V^*; \theta), \quad (9)$$

where  $l$  is the length of output word sequence  $Y = (y_1, \dots, y_l)$ .  $\theta$  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.

$$L = L_M + L_G. \quad (10)$$

## 4. Experiments

To evaluate the effectiveness of the proposed Visual Semantic Reasoning Network (VSRN), we perform experiments in terms of sentence retrieval (image query) and image retrieval (sentence query) on two publicly available datasets. Ablation studies are conducted to investigate each component of our model. We also compare with recent state-of-the-art methods on this task.

### 4.1. Datasets and Protocols

We evaluate our method on the Microsoft COCO dataset [28] and the Flickr30K dataset [39]. MS-COCO includes 123,287 images, and each image is annotated with 5 text descriptions. We follow the splits of [5, 8, 15, 23] for MSCOCO, which contains 113,287 images for training, 5,000 images for validation and 5000 images for testing. Each image comes with 5 captions. The final results are obtained by averaging the results from 5 folds of 1K test images or testing on the full 5K test images. Flickr30K consists of 31783 images collected from the Flickr website. Each image is accompanied with 5 human annotated text descriptions. We use the standard training, validation and testing splits [15], which contain 28,000 images, 1000 images and 1000 images respectively. For evaluation matrix, as is common in information retrieval, we measure the performance by recall at K (R@K) 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 set the word embedding size to 300 and the dimension of the joint embedding space  $D$  to 2048. We follow the same setting as [1, 23] to set details of visual bottom-up attention model. The order of regions for GRU-based global semantic reasoning (Sec. 3.3) is determined by the

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
(R-CNN, AlexNet)						
DVSA <sub>CVPR'15</sub> [15]	38.4	69.9	80.5	27.4	60.2	74.8
HMLstm <sub>ICCV'17</sub> [32]	43.9	-	87.8	36.1	-	86.7
(VGG)						
FV <sub>CVPR'15</sub> [20]	39.4	67.9	80.9	25.1	59.8	76.6
OEM <sub>ICLR'16</sub> [35]	46.7	-	88.9	37.9	-	85.9
VQA <sub>ECCV'16</sub> [29]	50.5	80.1	89.7	37.0	70.9	82.9
SMLstm <sub>CVPR'17</sub> [13]	53.2	83.1	91.5	40.7	75.8	87.4
2Way <sub>NCVPR'17</sub> [4]	55.8	75.2	-	39.7	63.3	-
(ResNet)						
RRF <sub>ICCV'17</sub> [30]	56.4	85.3	91.5	43.9	78.1	88.6
VSE++BMVC'18 [5]	64.6	89.1	95.7	52.0	83.1	92.0
GXN <sub>CVPR'18</sub> [8]	68.5	-	97.9	56.6	-	94.5
SCO <sub>CVPR'18</sub> [14]	69.9	92.9	97.5	56.7	87.5	94.8
(Faster R-CNN, ResNet)						
SCAN <sub>ECCV'18</sub> [23]	72.7	94.8	<b>98.4</b>	58.8	88.4	94.8
VSRN (ours)	<b>76.2</b>	<b>94.8</b>	98.2	<b>62.8</b>	<b>89.7</b>	<b>95.1</b>

Table 1. Quantitative evaluation results of the image-to-text (caption) retrieval and text-to-image (image) retrieval on MS-COCO 1K test set in terms of Recall@K (R@K).

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
(R-CNN, AlexNet)						
DVSA <sub>CVPR'15</sub> [15]	11.8	32.5	45.4	8.9	24.9	36.3
(VGG)						
FV <sub>CVPR'15</sub> [20]	17.3	39.0	50.2	10.8	28.3	40.1
VQA <sub>ECCV'16</sub> [29]	23.5	50.7	63.6	16.7	40.5	53.8
OEM <sub>ICLR'16</sub> [35]	23.3	-	84.7	31.7	-	74.6
(ResNet)						
VSE++BMVC'18 [5]	41.3	69.2	81.2	30.3	59.1	72.4
GXN <sub>CVPR'18</sub> [8]	42.0	-	84.7	31.7	-	74.6
SCO <sub>CVPR'18</sub> [14]	42.8	72.3	83.0	33.1	62.9	75.5
(Faster R-CNN, ResNet)						
SCAN <sub>ECCV'18</sub> [23]	50.4	<b>82.2</b>	<b>90.0</b>	38.6	69.3	80.4
VSRN (ours)	<b>53.0</b>	81.1	89.4	<b>40.5</b>	<b>70.6</b>	<b>81.1</b>

Table 2. 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).

descending order of their class detection confidence scores that are generated by the bottom-up attention detector. For the training of VSRN, we use the Adam optimizer [17] to train the model with 30 epochs. We start training with learning rate 0.00002 for 15 epochs, and then lower the learning rate to 0.000002 for the rest 15 epochs. We set the margin  $\alpha$  in Eq. 8 to 0.2 for all experiments. We use a mini-batch size of 128. For evaluation on the test set, we tackle over-fitting by choosing the snapshot of the model that performs best on the validation set. The best snapshot is selected based on the sum of the recalls on the validation set.

### 4.3. Comparisons With the State-of-the-art

**Results on MS-COCO.** Quantitative results on MS-COCO 1K test set are shown in Table 1, where the proposed VSRN outperforms recent methods with a large gap of R@1. Following the common protocol [5, 14, 23], the results are obtained by averaging over 5 folds of 1K test

images. When comparing with the current best method SCAN [23], we follow the same strategy [23] to combine results from two trained VSRN models by averaging their predicted similarity scores. Our VSRN improves 4.8% on caption retrieval (R@1) and 6.8% on image retrieval (R@1) relatively. In Table 2, we also report results on MS-COCO 5K test set by testing on the full 5K test images and their captions. From the table, we can observe that the overall results by all the methods are lower than the first protocol. It probably results from the existence of more distractors for a given query in such a larger target set. Among all methods, the proposed VSRN still achieves the best performance, which again demonstrates its effectiveness. It improves upon the current state-of-the-art, SCAN with 5.2% on the sentence retrieval (R@1) and 4.9% on the image retrieval (R@1) relatively.

**Results on Flickr30K.** We show experimental results of VSRN on Flickr30K dataset and comparisons with the current state-of-the-art methods in Table 3. We also list the network backbones used for visual feature extraction, such as R-CNN, VGG, ResNet, Faster R-CNN. From the results, we find the proposed VSRN outperforms all state-of-the-art methods, especially for Recall@1. When compared with SCAN [23] that uses the same feature extraction backbones with us, our VSRN improves 5.8% on caption retrieval (R@1) and 12.6% on image retrieval(R@1) relatively (following the same strategy [23] of averaging predicted similarity scores of two trained models). SCAN tries to discover the full latent alignments between possible pairs of regions and words, and builds up an attention model to focus on important alignments when inferring the image-text similarity. It mainly focuses on local pair-wise matching between regions and words. In contrast, the proposed VSRN performs 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. The comparison shows the strength of region relationship reasoning and global semantic reasoning for image-text matching. Especially for the challenging caption retrieval task, VSRN shows strong robustness to distractors with a huge improvement (relative 12.6%).

#### 4.4. Ablation Studies

**Analysis each reasoning component in VSRN.** We would like to incrementally validate each reasoning component in our VSRN by starting from a very basic baseline model which does not perform any reasoning. This baseline model adopts a mean-pooling operation on the region features after the fully-connected layer  $V = \{v_1, \dots, v_k\}$ ,  $v_i \in \mathbb{R}^D$  to obtain the final representation for the whole image  $I \in \mathbb{R}^D$ . The other parts are kept as the same as VSRN. Results on MS-COCO 1K test set are shown in Table 4.

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
(R-CNN, AlexNet)						
DVSA <sub>CVPR'15</sub> [15]	22.2	48.2	61.4	15.2	37.7	50.5
HMIstm <sub>ECCV'17</sub> [32]	38.1	-	76.5	27.7	-	68.8
(VGG)						
FV <sub>CVPR'15</sub> [20]	35.0	62.0	73.8	25.0	52.7	66.0
VQA <sub>ECCV'16</sub> [29]	33.9	62.5	74.5	24.9	52.6	64.8
SMIstm <sub>CVPR'17</sub> [13]	42.5	71.9	81.5	30.2	60.4	72.3
2WayN <sub>CVPR'17</sub> [4]	49.8	67.5	-	36.0	55.6	-
(ResNet)						
RRF <sub>ECCV'17</sub> [30]	47.6	77.4	87.1	35.4	68.3	79.9
VSE++BMVC'18 [5]	52.9	79.1	87.2	39.6	69.6	79.5
SCO <sub>CVPR'18</sub> [14]	55.5	82.0	89.3	41.1	70.5	80.1
(Faster R-CNN, ResNet)						
SCAN <sub>ECCV'18</sub> [23]	67.4	90.3	95.8	48.6	77.7	85.2
VSRN (ours)	<b>71.3</b>	<b>90.6</b>	<b>96.0</b>	<b>54.7</b>	<b>81.8</b>	<b>88.2</b>

Table 3. Quantitative evaluation 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).

This baseline model (noted as ‘Mean-pool’) achieves 64.3 of R@1 for caption retrieval and 49.2 of R@1 for image retrieval. Then we add one region relationship reasoning (RRR) layer (described in Sec. 3.3) before the mean-pooling operation into this baseline model and mark it as RRR. We also replace the mean-pooling operation with the global semantic reasoning (GSR) module (described in Sec. 3.3) to get a GSR model. From Table 4 we can find that these two reasoning modules can both help to obtain better image representation  $I$  and improve the matching performance effectively.

We then combine RRR and GSR to get our VSRN model and further try different numbers of RRR layers. Results show that adding region relationship reasoning layers before the global semantic reasoning module can gradually help to achieve better performance. This is because the RRR module can generate relationship enhanced features, which allows GSR perform reasoning on a graph topology and consider both current local region and global semantic correlations. However, we also find improvements become less when adding more RRR layers. We finally take 4RRR+GSR as the final setting of VSRN. We further report results of VSRN trained without text generation loss  $L_G$  (marked as 4RRR+GSR\*). Comparison shows that the joint optimization of matching and generation can help to improve around 2% relatively for R@1.

**Region ordering for global semantic reasoning.** Since our global semantic reasoning module (Sec. 3.3) sequentially processes region features and generates the representation of the whole image gradually, we consider several ablations about region ordering for this reasoning process in Table 5. 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

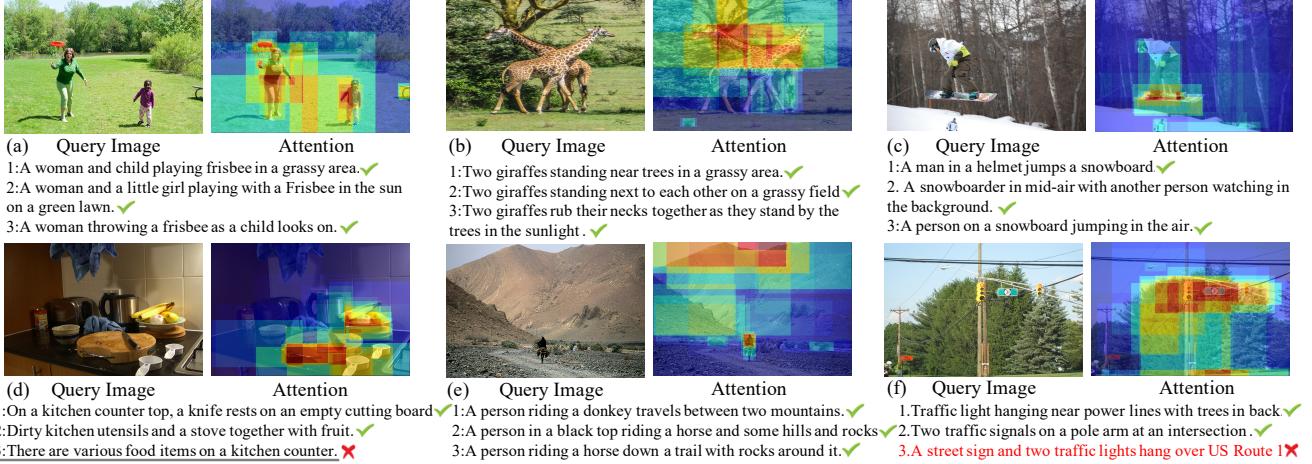


Figure 3. Qualitative results of the image-to-text (caption) retrieval for VSRN on MS-COCO dataset. For each image query, we show the top-3 ranked text caption. Ground-truth matched sentences are with check marks, while some sentences sharing similar meanings as ground-truth ones are marked with gray 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.)

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Mean-pool	64.3	90.5	95.1	49.2	83.4	91.5
RRR	68.5	93.2	96.3	56.8	87.2	94.2
GSR	72.3	94.4	98.0	59.6	88.6	94.5
1RRR + GSR	75.3	94.7	98.1	62.1	89.2	94.9
4RRR + GSR	76.2	94.8	98.2	62.8	89.7	95.1
4RRR + GSR*	74.6	94.6	98.2	61.2	89.0	94.8

Table 4. Ablation studies on the MS-COCO 1K test set. Results are reported in terms of Recall@K (R@K). “RRR” means model with region relationship reasoning module. “GSR” represents a model with global semantic reasoning module. The number before RRR represents the number of RRR layers. “\*\*” means model training without using text generation loss  $L_G$ .

Methods	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
VSRN-Random	75.1	94.5	98.0	62.3	89.1	94.6
VSRN-BboxSize	75.8	94.9	98.4	62.5	89.5	94.8
VSRN-Confidence	76.2	94.8	98.2	62.8	89.7	95.1

Table 5. Ablation studies on the MS-COCO 1K test set to analyze region ordering for global semantic reasoning. Results are reported in terms of Recall@K (R@K).

regions with high confidence first and then inferring more difficult regions based on the semantic context. 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 5 show that reasoning in a specific order can help improve the performance than the random one. VSRN-Confidence 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 included during the region relationship reasoning step. Based on these relationship enhanced feature, semantic reasoning can be then performed on global graph topologies.

#### 4.5. Visualization and Analysis

**Attention visualization of the final image representation.** Since the final goal of our visual semantic reasoning is to generate the image representation that includes key object and semantic concepts in the scene. In order to validate this, we visualize the correlation between the final representation of the whole image and these image regions included in this image in an attention form. Specifically, we calculate the inner product similarity (same as in Eq. 8) 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 score  $s_i$  to each  $v_i^*$  according to its rank  $r_i$ . The score is calculated by  $s_i = \lambda(k - r_{attn})^2$ , where  $k$  is the total number of regions,  $\lambda$  is a parameter used to emphasize the high ranked regions. We set  $\lambda = 50$  in our experiments. Then for the final attention map (similarity map), the attention score at each pixel location is obtained by adding up scores of all regions it belongs to. We show attention maps of each image along with the qualitative results of the image-to-text (caption) retrieval and text-to-image (image) retrieval.

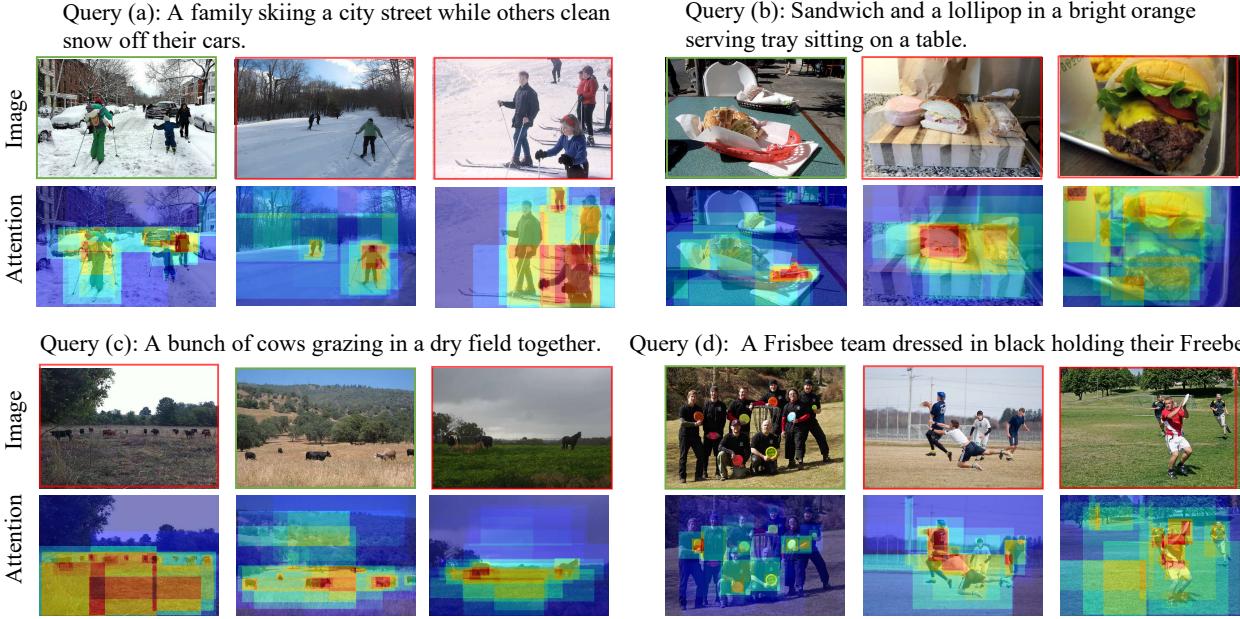


Figure 4. Qualitative results of the text-to-image (image) retrieval for VSRN on MS-COCO dataset. We show the top-3 retrieved images for each text query, ranking from left to right. The true matches are outlined in green boxes and false matches in red boxes. We also show the attention visualization of image representation generated by VSRN under the corresponding image.

**Qualitative results of the image-to-text retrieval.** In Figure 3, we show the qualitative results of the text retrieval given image queries on MS-COCO. We show the top-3 retrieved sentences for each image query. The rank is obtained based the similarity scores predicted by VSRN. From these results, we find that our VSRN can retrieve the correct results in the top ranked sentences even for cases of cluttered and complex scenes. The model outputs some reasonable mismatches, e.g. (d)-3. There are incorrect results such as (f.4), which is possibly due to the too specific concept in the image (“US Route 1”) that the model could not identify. From the attention visualization, we can find the image representation generated by VSRN well captures key objects and semantic concepts in the scene.

**Qualitative results of the text-to-image retrieval.** In Figure 4, we show qualitative results of image retrieval for given text queries on MS-COCO. Each text description is matched with a ground-truth image. We show the top-3 retrieved images for each text query. The true matches are outlined in green and false matches in red. We find that our model can retrieve the ground truth image in the top-3 list. Note that other results are also reasonable, which include the objects of the same category or same semantic concepts with the text descriptions. For those images with a very similar scene, our model can still distinguish them well and accurately retrieve the ground truth one at top-1 rank. This can be well explained from the attention map, e.g. for the given text query (a), the model attends on the cars on the street and the person cleaning a car in the ground-truth im-

age to distinguish it with the other two images that are also about people skiing. However, for the top-2 retrieval images of query (c), the model is confused about the concept of “try field”. It treats the field with less grass as a better match than the field with withered grass. This is possibly due to not enough training data for a complex concept.

## 5. Conclusion

In this paper, we present a simple and interpretable reasoning model VSRN to generate visual representation by region relationship reasoning and global semantic reasoning. The enhanced image representation captures key objects and semantic concepts of a scene, so that it can better align with the corresponding text caption. Extensive experiments on MS-COCO and Flickr30K datasets demonstrate the resulting model consistently outperforms the-state-of-the-art methods with a large margin for the image-text matching. Compared with the complicated attention-based aggregation from pairwise similarities among regions and words, we show that the classical “image-text” similarity measure still promising given enhanced whole image representation. We will further explore the effectiveness of reasoning modules in VSRN on other vision and language tasks.

## 6. Acknowledgments

This research is supported in part by the NSF IIS award 1651902 and U.S. Army Research Office Award W911NF-17-1-0367.

## References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [2] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *CVPR*, 2018.
- [3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*, 2014.
- [4] Aviv Eisenshtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.
- [5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.
- [6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.
- [7] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [8] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, 2018.
- [9] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [11] Jerry R Hobbs, Mark E Stickel, and Paul Martin. Interpretation as abduction. *Artificial intelligence*, 1993.
- [12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.
- [13] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *CVPR*, 2017.
- [14] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018.
- [15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [16] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 2014.
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.
- [18] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [19] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014.
- [20] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.
- [21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.
- [22] Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *EMNLP*, 2011.
- [23] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.
- [24] Kai Li, Zhengming Ding, Kunpeng Li, Yulun Zhang, and Yun Fu. Support neighbor loss for person re-identification. In *ACM Multimedia*, 2018.
- [25] Kunpeng Li, Ziyuan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018.
- [26] Kunpeng Li, Ziyuan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *TPAMI*, 2019.
- [27] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Attention bridging network for knowledge transfer. In *ICCV*, 2019.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [29] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.
- [30] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.
- [31] Allen Newell. Physical symbol systems. *Cognitive science*, 1980.
- [32] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *ICCV*, 2017.
- [33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [34] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.
- [35] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.
- [36] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, 2015.
- [37] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *ICLR*, 2019.
- [38] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.
- [39] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.
- [40] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *ICCV*, 2019.
- [41] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.