Know More Say Less: Image Captioning Based on Scene Graphs

Xiangyang Li and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Automatically describing the content of an image has been attracting considerable research attention in the multimedia field. To represent the content of an image, many approaches directly utilize convolutional neural networks (CNNs) to extract visual representations, which are fed into recurrent neural networks to generate natural language. Recently, some approaches have detected semantic concepts from images and then encoded them into high-level representations. Although substantial progress has been achieved, most of the previous methods treat entities in images individually, thus lacking structured information that provides important cues for image captioning. In this paper, we propose a framework based on scene graphs for image captioning. Scene graphs contain abundant structured information because they not only depict object entities in images but also present pairwise relationships. To leverage both visual features and semantic knowledge in structured scene graphs, we extract CNN features from the bounding box offsets of object entities for visual representations, and extract semantic relationship features from triples (e.g., *man riding bike*) for semantic representations. After obtaining these features, we introduce a hierarchical-attention-based module to learn discriminative features for word generation at each time step. The experimental results on benchmark datasets demonstrate the superiority of our method compared with several state-of-the-art methods.**

***Index Terms*—Image captioning, scene graph, relationship, long short-term network, attention mechanism, vision-language.**

I. Introduction

IMAGE captioning is the task of automatically describing the content of images with natural language sentences, which has received increasing attention in the field of multimedia and artificial intelligence [1]–[7]. This task not only connects computer vision and natural language processing which are two prominent ways to acquire knowledge, but also has many applications such as semantic image search [8]–[10], vision-and-language navigation [11]–[13], injecting visual intelligence into chatbots, and helping visually impaired people to perceive the visual world around them. Unlike image classification [14]–[16] and object detection [17]–[19] which assign predefined labels for images, image captioning aims to describe images with free-form natural language.

Over the past few years, remarkable progress has been achieved in image captioning. Most of the prominent approaches [3], [20]–[25] are based on the encoder-decoder framework [26]. For example, Vinyals *et al.* [20] first extract image features from the last fully connected layer of a convolutional neural network (CNN), and then feed the features into a long short-term memory (LSTM) network to generate descriptions. In order to reveal more details in images, different from [20], Xu *et al.* [21] use feature maps from the last convolutional layer from a pre-trained CNN to represent the image and then use attention mechanism on spatial locations to obtain more representative features. Jin *et al.* [22] utilize attention mechanism on detected objects when generating language descriptions. Although prominent performance can be achieved by these methods, they directly translate visual features into sentences, ignoring high-level semantic features that provide a deeper understanding of images and that contain useful information for image captioning. To address this issue, many methods explicitly exploit semantic features for images [3], [5], [23], [27]–[29]. For example, Wu *et al.* [27] treat the semantic features as an initial word for the decoder. You *et al.* [23] employ the detected semantic concepts in images for text generation.

The existing approaches for image captioning treat all the identified entities (e.g., objects, semantic concepts) individually without considering relationships among them. Note that in this paper, we refer to the interaction between two objects as a relationship (e.g., riding, under, and next to, as shown in Fig. 1). However, relationships are the main linguistic components for captionsbecause captions often describe multiple objects in images. For example, as illustrated in the top left of Fig. 1, we prefer to describe the image as "a boy lying prone on a surfboard" rather than as " a boy and a surfboard". We can note that relationships often appear in image descriptions, which is consistent with our habitual expression ways, as we often describe objects with their relationships rather than describing them individually. The statistical information in the VG-COCO dataset (which will be introduced in Section IV-A) also corresponds to this observation. As shown in Fig. 2, ground-truth relationships with high frequencies in images also appear with high frequencies in ground-truth captions.

Manuscript received July 3, 2018; revised October 31, 2018 and December 31, 2018; accepted January 2, 2019. Date of publication January 30, 2019; date of current version July 19, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61532018, in part by the Beijing Natural Science Foundation under Grant L182054, in part by the National Program for Special Support of Eminent Professionals, and in part by the National Program for Support of Top-notch Young Professionals. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Engin Erzin. *(Corresponding author: Shuqiang Jiang.)*

The authors are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xiangyang.li@vipl.ict.ac.cn; sqjiang@ict.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2896516

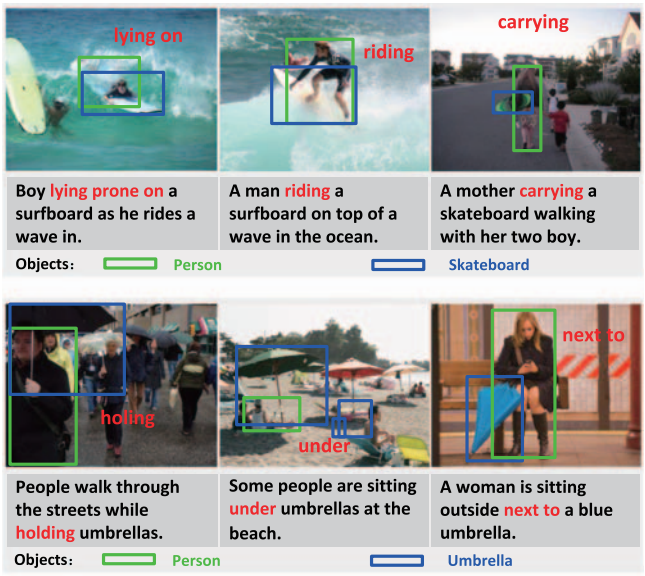

Fig. 1. Examples of images that have the same compositional objects but different relationships (e.g., holding, next to, and under). Different relationships bring different holistic interpretations of the images. Text under the image is the corresponding ground-truth caption (Best viewed in color).

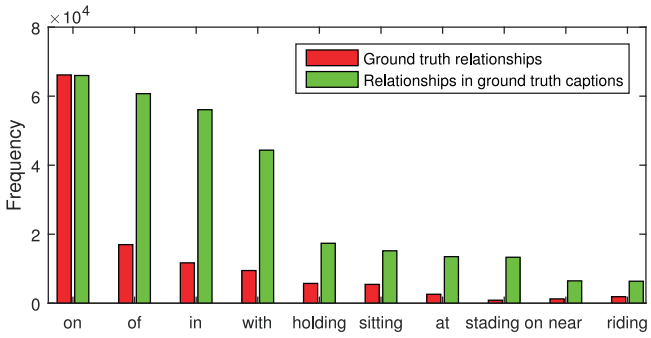

Fig. 2. The ground-truth (GT) relationships and the ones in GT captions. This figure shows that relationships with high frequencies in the images appear with high frequencies in the GT captions. Because each image has 5 corresponding sentences, the number of relationships in the captions is larger than the number of GT relationships. For simplicity, we only show the top 10 relationships.

Relationships are of crucial importance for describing the content of images. First, relationships reveal more information contained in images compared to image classification and object detection. Typically, a relationship is a structured knowledge that inherently bundles two objects together. Therefore, it is a kind of middle representation that connects local regions and the global image. For example, as shown in the bottom row of Fig. 1, given an image with persons and umbrellas, localizing and recognizing individual objects in the image are not enough to generate an appropriate description, as the persons in the image may holding, next to, or under umbrellas. Reasoning the relationships between objects can provide an in-depth understanding of the image thus forming the holistic interpretations. Second, because relationships and their corresponding objects are often skeletons of captions, their mutual information can avoid generating unreasonable captions. For example, as shown

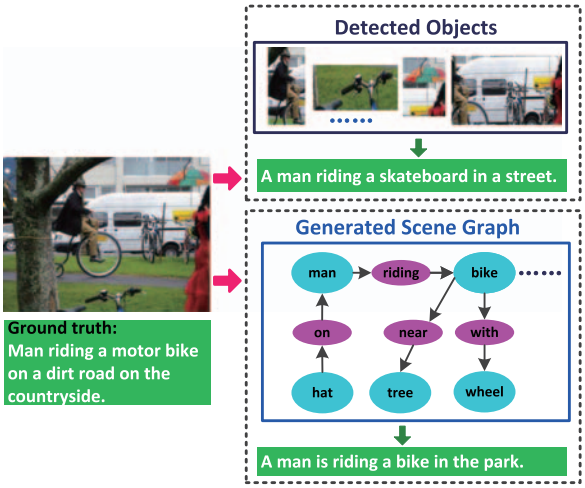

Fig. 3. Illustration of two image captioning schemes. The first row shows an example generated by a baseline method that utilizes detected objects to generate descriptions. The second row shows a more accurate description generated by our method based on scene graphs that capture objects (blue nodes) and their relationships (red edges).

in Fig. 3, only using the objects in the image, the baseline method generates an improper caption (i.e., *a man riding a skateboard in a street*), which falsely predicts the bike as a skateboard. With the generated sequence "a man riding a", it appearsto be reasonable for the baseline method to generate the word "skateboard" because there are many examples of "man riding skateboard" in the training data. With the detected triples such as "man riding bike", our method which incorporates relationships in the image is able to generate a more accurate caption (i.e., a *man is riding a bike in the park*).

Whereas relationships are representations of local image regions, scene graphs are structured representations for the whole images [30]–[32]. In such graphs, the nodes correspond to object bounding boxes with their object categories, and edges correspond to their pairwise relationships between objects. They contain rich relationships that pinpoint different aspects of images. However, because a caption often describes parts of objects in an image, different parts of a scene graph have different relevance for generating a description. For example, as shown in Fig. 3, the part which is composed of the relationship edge of *riding* and its two object nodes (i.e., *man* and *bike*) are the most relevant for generating the caption. Therefore, it is important to select relevant information from scene graphs when describing images.

In this paper, we propose a novel framework which is based on scene graphs to generate descriptions for images, as shown in Fig. 4. To the best of our knowledge, this work is the first attempt to explicitly employs knowledge in scene graphs for image captioning. We first extract CNN features of the corresponding regions of object entities for visual representations. Because relationships depict different aspects of images, we extract triples (e.g., *man riding bike*) which are three lexeme sequences from scene graphs. These triples are then embedded into semantic vectors, forming the semantic relationship features. After obtaining these features, we introduce a

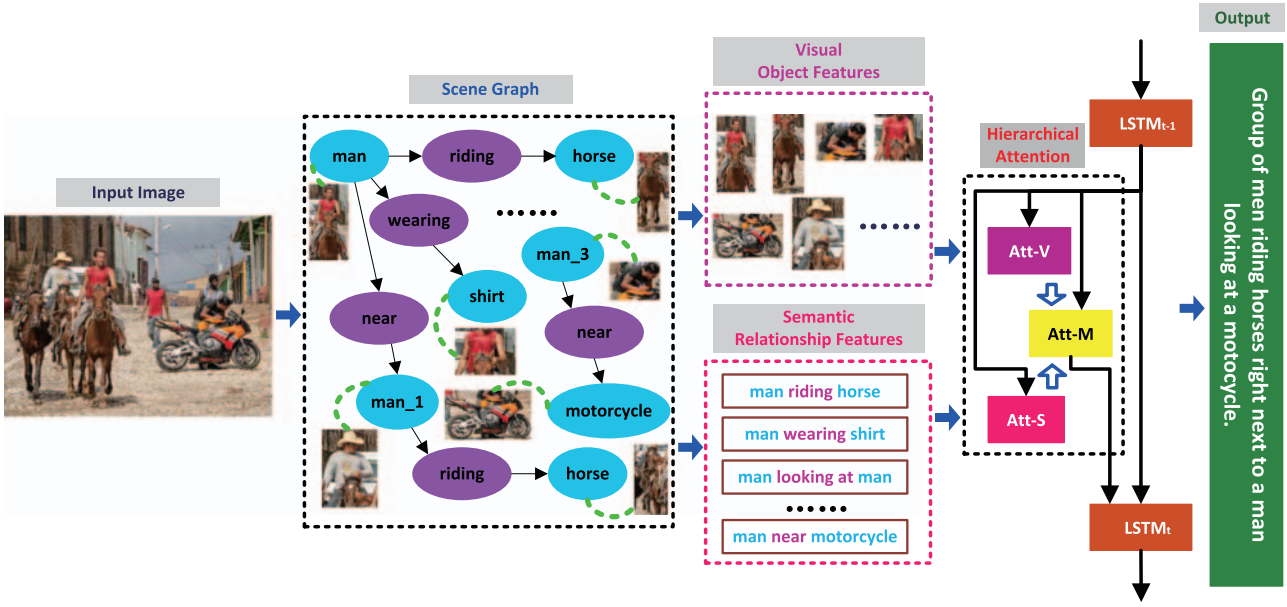

Fig. 4. The framework of our proposed method. First, the scene graph for the input image is generated. To obtain visual features, we extract CNN features from the corresponding regions of object entities. To obtain semantic features, we extract triples, which are three lexeme sequences that describe object relationships from the graph, and embed them into fix-length vectors. To leverage both kinds of these information, a hierarchical-attention-based fusion module is introduced to decide when and what to attend to during the sentence generation process.

hierarchical-attention-based module to automatically learn key cues for word generation, as a caption is only related to some of the rich information that is contained in an image. With the first-level attention, it selectively attends to different visual features and semantic features, forming the weighted visual and semantic context vectors, respectively. Then it learns relevance scores for these two modalities to integrate them with the second-level attention. Our approach achieves promising results compared with state-of-the-art methods.

In summary, our contributions are as follows:

- We propose an architecture based on scene graphs for generating natural language descriptions. Because scene graphs not only depict object entities but also present pairwise relationships, they provide abundant information for describing images.
- We extract visual features from object entities and semantic relationship features from triples extracted from scene graphs. Additionally, we introduce a hierarchical-attention-based module that adaptively attend to these features during word generation. The results demonstrate the effectiveness of our proposed method.

The remainder of this paper is organized as follows. We provide a brief overview of related work in Section II. In Section III, we describe our method that explicitly utilizes the relationships in scene graphs for image captioning. We then present the quantitative and qualitative results in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

Because this work is primarily related to the topics of image captioning, visual relationship detection, scene graphs and attention mechanism, we briefly review the most recent literature on these approaches.

A. Image Captioning

Image captioning is a complicated cognitive task that bridges the gap between image understanding and natural language processing. Some early approaches first retrieve images from a corpus with image-caption pairs and then directly transfer the corresponding human-written sentences for the query images [33], [34]. Other pioneering methods are dedicated to exploring bottom-up schemes that are based on templates. These methods first detect elements with descriptive information (e.g., nouns, adjectives, verbs) from images and then use predefined sentence templates to generate descriptions [35], [36]. Recently, prevailing methods based on neural networks have adopted convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to address this task [3]–[6], [20], [22], [24], [28], [37]–[39]. Karpathy *et al.* [40] and Mao *et al.* [41] use a deep CNN to extract visual features and put these features into an RNN as the initial start word to generate image descriptions. Vinyals *et al.* [21] start to employ a long short-term memory (LSTM) network to model language as it includes a memory cell that can maintain information in memory for long periods of time. Jia *et al.* [29] propose gLSTM which is guided with extra semantic information extracted from images. Li *et al.* [42] utilize both object and scene information in images. To address the issue that LSTM units are complex and inherently sequential across time, Aneja *et al.* [43] propose a convolutional image captioning approach, which operates over all words in parallel. In order to generate natural language explicitly grounded in the entities found by object detectors, Lu *et al.* [44] propose a two-stage

approach that first generates a hybrid template and then fills the slots in the template based on categories recognized by object detectors. Rennie *et al.* [45] and Liu *et al.* [46] utilize a reinforcement learning (RL) approach to optimize their language model. Lin *et al.* [47] modify their language model by integrating a generative adversarial Network (GAN) for generating high-quality language descriptions. Although most of these approaches aim to improve language models with more complicated submodules or enhancing the training procedure, our work is orthogonal to these methods. In this paper, we explore more comprehensive image representations for image captioning. Similar to our work, Chen *et al.* [5] obtain structured representations of images based on simplified visual parsing trees with a fixed structure, whereas our representations of images are more comprehensive structured scene graphs.

B. Visual Relationship Detection

As the intermediate-level task that connects image captioning and object detection, visual relationship detection (VRD) not only recognizes objects in an image but also predicts their relationships. Earlier approaches explore specific types of relationships (e.g., spatial relations such as "below", "above") and often utilize them to facilitate other tasks [48], [49]. In recent years, new methods have been developed for generic VRD [50]–[53]. Lu *et al.* [50] train a visual appearance module as well as a prior language module and then combine them to predict multiple relationships per image. Due to the massive semantic space of visual relationships, Yu *et al.* [52] obtain linguistic knowledge by mining from both training annotations and publicly available text to regularize visual model learning. Li *et al.* [54] propose a phrase-guided message passing structure to model the visual interdependence by refining features among subject, predicate and object branches. Dai *et al.* [55] propose deep relational networks to model the statistical dependencies between objects and their relationships. Zhang *et al.* [56] propose a visual translation embedding network for simultaneous object detection and relation prediction. Liang *et al.* [51] apply the reinforcement learning method to sequentially discover object relationships and attributes in an image. Because relationships are middle-level representations that connect local regions and the global understanding of images, in this paper, we explicitly employ semantic relationship features for image captioning.

C. Scene Graphs

Beyond the visual relationship detection task that detects multiple local relationships for an image, the scene graph generation task [30], [31], [57], [58] endows the image with an entire structured representation capturing both objects and their semantic relationships, where the nodes are object instances in images and the edges depict their pairwise relationships. Xu *et al.* [30] propose a novel model that generates such a structured scene representation from an input image by iterative message passing. Li *et al.* [31] leverage mutual connections across different semantic levels of image understanding to help generate scene graphs. Zellers *et al.* [57] analyze the role of motifs (i.e., regularly appearing substructures in scene graphs) in the Visual Genome dataset [59] and introduce strong baselines which model these intragraph interactions. Klawonn *et al.* [60] propose an approach which is based on generating individual subgraphs called triples and exploiting an attention mechanism to stitch triples together into a proper scene graph. Wang *et al.* [58] introduce a new perspective and solution for the task of generating scene graphs from textual descriptions. They first align nodes in region graphs with words in the region descriptions using simple rules, and then they use this alignment to train their customized dependency parser. In our work, to obtain structured representations from visual images, we use the approach proposed by Xu *et al.* [30] to generate scene graphs for images.

Scene graph representations have been proven to be useful in various visual tasks [32], [61], [62]. Johnson *et al.* [32] use ground-truth scene graphs that are generated by humans and grounded to real-world images as queries to retrieve semantically related images. Teny *et al.* [61] propose building graphs over the scene objects and over the question words, and then they describe a deep neural network that exploits the structure in these representations for visual question answering. However, their experiments are implemented on a dataset of clip art images, in which the corresponding graphs of the images are provided. In our work, we perform experiments on real-word images in which scene graphs are learned automatically. Johnson *et al.* [62] propose a method for generating images from scene graphs which are obtained from textual descriptions. In this work, we utilize scene graphs which are generated from visual images to generate natural language descriptions.

D. Attention Mechanism

The attention mechanism is widely known in psychology and neuroscience. This mechanism has recently been introduced in machine translation [63] to generate weights of the individual words of the sentence to be translated. Rather than encoding the input sequence into a single fixed context vector, it allows the decoder to attend to different parts of the source sentence at each step of the output generation. Xu *et al.* [21] first propose a spatial attention model over feature maps of images to generate sentences, in which the LSTM units update with the weighted visual representation of image regions. In their work, the attention weight for each region is determined based on the previous state of the LSTM. Liu *et al.* [64] take a step further by proposing a quantitative method to evaluate the consistency between the generated attention maps and ground-truth annotated regions. In their work, they also add explicit supervision for the attention map learning procedure. Jin *et al.* [22] apply attention on detected visual objects when describing images. To further improve performance, You *et al.* [23] use attention on rich high-level semantic concepts which are detected from images. Yang *et al.* [65] add a reviewer module that improves the representation passed to the decoder. Xu *et al.* [66] use the attention mechanism to obtain useful context for video captioning. In our work, we propose a hierarchical-attention-based module to automatically select visual features and semantic

relationship features from scene graphs during the generation of image captions.

III. Our Approach

The overview of our image captioning pipeline is illustrated in Fig. 4. To obtain structured representations of images, the scene graphs of these images are generated. To take full use of information contained in scene graphs, we extract both visual features and semantic relationship features from scene graphs. Additionally, we use a fusion module based on hierarchical attention to automatically select these features for word generation. In this section, we first describe the generic encoder-decoder framework for image captioning, and then we introduce our approach which explicitly leverages the information contained in scene graphs for generating natural language descriptions of images.

A. Encoder-Decoder for Image Captioning

In this subsection, we briefly describe the encoder-decoder image captioning framework [20], [21], [23]. Image captioning takes an image I as the input, and generates a natural language sentence s.

$$G:\ I \rightarrow s \tag{1}$$

In order to solve this problem, we design a model $P_G(s|I)$. With P_G, we have:

$$G(I) = \underset{s}{argmax}\ P_G(s|I) \tag{2}$$

To optimize the parameters in P_G, we need a dataset with a set of image-sentence pairs $\{(I_i s_i)\}$. Since we often use an encoder-decoder based model to represent $P_G(s|I)$, where the core idea is generally to maximize the probability of the description given the input image, we can generate the terms of s one after another until the end of it. The entire model is trained to minimize cross-entropy loss which is equivalent to maximizing the likelihood:

$$L_{loss} = -\sum_{i}^{N}\sum_{t=1}^{T}(logP_G(s_{i,t}|s_{i,\ 1:t-1},\ I_i)) \tag{3}$$

where N is the total number of samples in the training, $s_{i,t}$ is the t-th word of the ground-truth caption s_i, and T is the length of s_i.

In the encoder-decoder framework, a recurrent neural networks (RNNs) is often utilized as the decoder to generate captions. In this paper, we adopt a long short-term memory (LSTM) [67] as our language model. It takes the output of the previous time step, the input of the current time step and the context extracted from image I as its current inputs. The update equations at time t are formulated as:

$$i_t = \sigma(\mathbf{W}_{ix}x_t + \mathbf{W}_{im}h_{t-1} + \mathbf{W}_{ig}k_t) \tag{4}$$

$$f_t = \sigma(\mathbf{W}_{fx}x_t + \mathbf{W}_{fm}h_{t-1} + \mathbf{W}_{fg}k_t) \tag{5}$$

$$\mathrm{o}_t = \sigma(\mathbf{W}_{ox}x_t + \mathbf{W}_{om}h_{t-1} + \mathbf{W}_{og}k_t) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(\mathbf{W}_{cx}x_t + \mathbf{W}_{cm}h_{t-1} + \mathbf{W}_{cg}k_t) \tag{7}$$

$$h_t = \mathrm{o}_t \odot \phi(c_t) \tag{8}$$

where t ranges from the start of the input sequence to the end of it; i_t, f_t and o_t represent the input gate, forget gate, and output gate at time step t, respectively; c_t is the state of the memory cell and h_t is the hidden state; k_t is the context at time t; x_t is the element of the sequence at timestep t; $\odot$ represents the element-wise multiplication, $\sigma(\cdot)$ represents the sigmoid function, and $\phi(\cdot)$ represents the hyperbolic tangent function; and $\mathbf{W}_{[\cdot][\cdot]}$ denote the parameters of the model. For simplicity, in the next of our paper, the update procedure is simplified as:

$$h_t = LSTM(x_t,\ h_t,\ k_t) \tag{9}$$

The context k_t is an important factor in the encoder-decoder framework, and it provides rich evidence for generating descriptions [21]–[23], [29], [64]. The approaches which model the context vector can be divided into two categories. The first one is fixed context which is extracted from images [29], [41]. During word generation, the kind of context remains constant and does not depend on the decoder. The second one is variable context which depends on the hidden state of the decoder [21], [23]. At each time step t, specific elements are selected from a set of context vectors, thus forming discriminative context which helps improve the performance of image captioning. In this work, we use an attention-based approach to obtain useful context from scene graphs. In the next subsections, we first introduce how we generate scene graphs from images and then describe the way we exploit both visual and semantic features from scene graphs for image captioning.

B. Scene Graph Generation and Feature Extraction

Scene graphs not only depict object entities in images (with bounding boxes and categories) but also present pairwise relationships. They are structured representations which contain rich relationships, providing abundant information for describing images. In order to generate a visually grounded scene graph that most accurately correlates with an image, we use the method proposed by Xu *et al.* [30] to generate scene graphs for images, which is based on recurrent neural networks to predict objects and their relationships.

To generate a scene graph for an image, a set of initial bounding boxes should first be generated. In our paper, we use the region proposal network (PRN) proposed by Girshick *et al.* [17] to generate a set of object proposals for the image I. Second, for each proposal, its object category as well as its bounding box offsets need to be predicted. What is more, the relationship between each pair of objects also needs to be considered. We formulate the scene generation problem as follows. Given a set of classes $\mathcal{C}$ (including the background class) and a set of relationship types $\mathcal{R}$ (including the none relationship), the variables that need to be predicted in a scene graph are denoted as:

$$P = \left\{p_i^{cls},\ p_i^{bbox},\ p_{i\rightarrow j}|i = 1, ..., n;\ j = 1, ..., n;\ i \neq j\right\} \tag{10}$$

where n is the number of proposal boxes, $p_i^{cls} \in \mathcal{C}$ is the class label for the i–th proposal box, $p_i^{bbox} \in \mathbb{R}^4$ is the bounding box offsets of the i–th proposal box, and $p_{i\rightarrow j} \in \mathcal{R}$ is the relationship predicate between the i–th and j–th proposal boxes. At the holistic level, the inference task is to find the

optimal $P^* = \underset{P}{argmax} Pr(P|I,\ B_I)$ that maximizes the following probability function:

$$Pr(P|I,\ B_I) = \prod_{i}\prod_{j\neq i} Pr(p_i^{cls},\ p_i^{bbox},\ p_{i\rightarrow j}|I,\ B_I) \quad (11)$$

Too employ contextual information among objects, the inference procedure is based on RNNs, learning to interactively improve its predictions via message passing. The probability of each variable g (i.e., objects or relationships) in a scene graph is denoted as $Q(g|\cdot)$, assuming that the probability only depends on the current state of each node and each edge at every iteration. Gated recurrent units [68] are used to compute the hidden states of the variables due to their simplicity and effectiveness. A gated recurrent unit is a simplification of the LSTM architecture. The update equations of a GRU at time t are formulated as:

$$z_t^g = \sigma(\mathbf{W}_{zx} x_t^g + \mathbf{W}_{zh} h_{t-1}^g) \quad (12)$$

$$r_t^g = \sigma(\mathbf{W}_{rx} x_t^g + \mathbf{W}_{rh} h_{t-1}^g) \quad (13)$$

$$\tilde{h}_t^g = \phi(\mathbf{W}_{vx} x_t^g + \mathbf{W}_{vh}(r_t^g \odot h_{t-1}^g)) \quad (14)$$

$$h_t^g = (1 - z_t^g) h_{t-1}^g + z_t^g \tilde{h}_t^g \quad (15)$$

where z_t^g is the update gate which decides how much the unit updates its content; r_t^g is the reset gate; and h_t^g is the hidden state at time t, which is a linear interpolation between the previous hidden state h_{t-1}^g and the candidate hidden state $\tilde{h}_t^g$.

The current hidden state of node i is denoted as h_i^g and the current hidden state of edge $i \rightarrow j$ is denoted as $h_{i\rightarrow j}^g$. Then the inference procedure can be formulated as:

$$Q(P|I,\ B_I) = \prod_{i=1}^{n} Q(p_i^{cls},\ p_i^{bbox}|h_i^g) Q(h_i^g|f_i^v) \prod_{j\neq i} Q(p_{i\rightarrow j}|h_{i\rightarrow j}^g) Q(h_{i\rightarrow j}^g|f_{i\rightarrow j}^e) \quad (16)$$

where f_i^v is the feature vector of the i–th node, and $f_{i\rightarrow j}^e$ is the feature vector of the edge connecting the i–th node and the j–th node. In the first iteration, the GRU units take the visual features f_i^v (i.e., the visual features of the proposal box i) and $f_{i\rightarrow j}^e$ (i.e., the visual features of the union box over the proposal boxes i and j) as input. In later iterations, the inputs are the aggregated messages from other GRU units of the previous step. Details can be found in [30]. Based on h_i^g and $h_{i\rightarrow j}^g$, a softmax layer is utilized to produce the final scores for the object class and relationship predicate. Meanwhile, a fully connected layer is also used to regress to the bounding box offsets for each object class separately.

After obtaining scene graphs for images, we then extract representative and discriminative features based on them for generating image descriptions. Because visual features and semantic features describe image content at different levels, we extract low-level visual features as well as high-level semantic features for each image, leveraging information from both objects and their relationships. On the one hand, because objects are the core building blocks of images, we extract objects with high prediction scores from the generated scene graphs. Given an object o_i of an image I, we extract features from CNNs to represent it. In this manner, we can obtain the visual features O_I for image I, where $O_I = \{o_1,\ o_2,\ ...,\ o_M\}$. On the other hand, because a relationship inherently bundle two objects together and provides an in-depth understanding of the image, we extract triples which are three lexeme sequences (e.g., *man riding bike*, *man holding umbrella*) from scene graphs.

The reasons why we use such triples to represent knowledge in scene graphs arise from the following three aspects: 1) To extract information in graph-structured data, a triple which corresponds to two entities and an edge is generally treated as the basic element. For example, Borde *et al.* [69] sample such triples to reason the information contained in knowledge graphs. Klawonn *et al.* [60] stitch such triples together to form the entire scene graph for an image. 2) A relationship (i.e., a triple) is a small subgraph of the entire scene graph. It describes a single statement about a scene from a specific region. Meanwhile, a caption often describes parts of objects in the image, thus, different parts of the image have different relevances for generating the description. Therefore, a triple in a scene graph provides discriminative and informative cues for generating the caption. 3) Although individual predicates of relationships can provide some kind of semantic information, they are somewhat ambiguous. For example, given the predicate riding alone, we cannot exactly imagine what kind of visual content is contained, as a person can ride a bike or a horse. To encode a triple r_i, similar to previous approaches that represent a sentence or a phrase [29], [70], we utilize word2vec [71] to obtain the 300-dimensional vector for each word and then use mean pooling to obtain the integral one. Then the semantic relationship features R_I of an image can be obtained, where $R_I = [r_1,\ r_2,\ ...\ r_N]$.

C. Hierarchical-Attention-Based Feature Fusion

Assuming that we have obtained the visual features O_I and the semantic relationship features R_I of an image I, in this section, we propose a hierarchical-attention-based module to handle the fusion of these features, which can selectively attend to specific features that are relevant to predicting each subsequent word. The architecture of our feature fusion module is shown in Fig. 5.

Although a scene graph provides a full structured representation of an image, in many cases, a caption is only related to some of the information which is contained in an image. For example, in Fig. 3, the key clues of the caption are only the man and the bike. Using the global information to describe the image could lead to suboptimal results due to the noises introduced from information which is irrelevant to the potential caption. Thus, it is important to automatically learn important clues from scene graphs. With the first-level attention, our hierarchical-attention-based fusion module first selectively attends to different visual features and semantic features, forming the weighted visual and semantic context vectors, respectively. Rather than simply pooling these two kinds of context vectors into a single vector, which neglects the inherent structure and differences among them, it then learns relevance scores for these two modalities

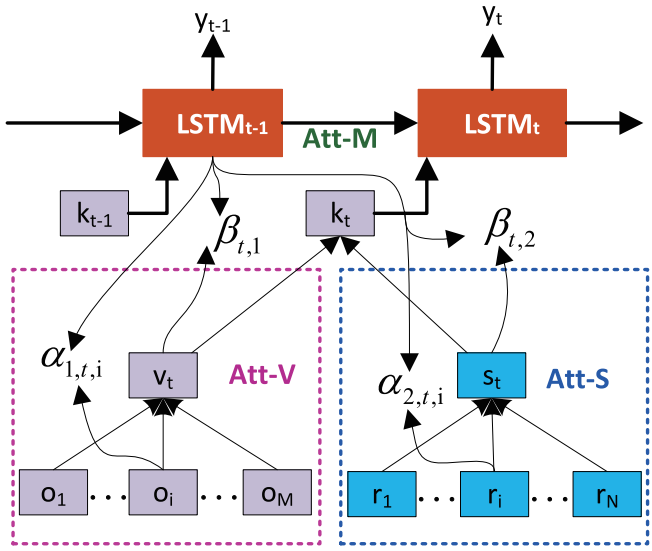

Fig. 5. Illustration of the hierarchical-attention-based feature fusion module which can selectively attend to specific features that are relevant for predicting the target word. It is composed of visual attention (Att-V), semantic attention (Att-S), and multimodal attention (Att-M).

to obtain the final context vector with the second-level attention.

To obtain the weighted visual context vector v_t at each time step t, we use visual attention (Att-V) to generate a normalized attention weight $\alpha_{1,t,i}$ for each of the M visual features o_i as follows:

$$\mathrm{a}_{1,t,i} = \mathcal{W}_o^T tanh(\mathbf{W}_{ov} o_i + \mathbf{W}_{og} h_{t-1}) \tag{17}$$

$$\alpha_{1,t,i} = \frac{exp(\mathrm{a}_{1,t,i})}{\sum_i exp(\mathrm{a}_{1,t,i})} \tag{18}$$

$$v_t = \sum_{i=1}^{M} \alpha_{1,t,i} o_i \tag{19}$$

where h_{t-1} is the previous hidden state of the decoding LSTM (from Eq. 9); $\mathbf{W}_{ov}$, $\mathbf{W}_{og}$and $\mathcal{W}_o$ represent the Att-V parameters that are estimated together with all other parameters in the decoding procedure. At the same time, to obtain the weighted semantic context vector s_t, we employ semantic attention (Att-S) to learn the relevance score $\alpha_{2,t,i}$ for each relationship of the N semantic relationship features, as shown in Fig. 5. The update procedure of Att-S is:

$$\mathrm{a}_{2,t,i} = \mathcal{W}_r^T tanh(\mathbf{W}_{rv} r_i + \mathbf{W}_{rg} h_{t-1}) \tag{20}$$

$$\alpha_{2,t,i} = \frac{exp(\mathrm{a}_{2,t,i})}{\sum_i exp(\mathrm{a}_{2,t,i})} \tag{21}$$

$$s_t = \sum_{i=1}^{N} \alpha_{2,t,i} r_i \tag{22}$$

where $\mathbf{W}_{rv}$, $\mathbf{W}_{rg}$ and $\mathcal{W}_r^T$ represent the Att-S parameters. After obtaining v_t and s_t, in the same spirit, we use multimodal attention (Att-M) which learns relevance scores $\beta_{t,1}$ for v_t and $\beta_{t,2}$ for s_t to obtain context from both visual and semantic modalities. The update procedure of Att-M is :

$$\mathrm{b}_{t,1} = \mathcal{W}_m^T tanh(\mathbf{W}_{mv} v_t + \mathbf{W}_{mg} h_{t-1}),$$

$$\mathrm{b}_{t,2} = \mathcal{W}_m^T tanh(\mathbf{W}_{mv} s_t + \mathbf{W}_{mg} h_{t-1}) \tag{23}$$

$$\beta_{t,i} = \frac{exp(\mathrm{b}_{t,i})}{\sum_i exp(\mathrm{b}_{t,i})} \tag{24}$$

$$k_t = \beta_{t,1} v_t + \beta_{t,2} s_t \tag{25}$$

where $\mathbf{W}_{mv}$, $\mathbf{W}_{mg}$ and $\mathcal{W}_m^T$ are the learned parameters of Att-M. The final context vector k_t will be fed into the decoder to generate captions.

D. Implementation Details

To obtain scene graphs for images, we train the scene graph model on the Visual Genome dataset [59] (except the images in the validation set of MS COCO [72]). We use the relationships which are manually cleaned by Xu *et al.* [30]. In this way, our scene graph model contains 150 object categories (excluding the background category) and 50 relationship types (excluding the none relationship). With the same settings asf [30], we also use the VGG-16 [15] network to extract features from images. To obtain scene graphs, the training procedure is almost the same with [30]. For training the language model, we use an LSTM network with a hidden state size of 512. We remove the words that occur less than 5 times in the training and validation sets, resulting dictionaries of size 9567 for MS COCO and 5551 for VG-COCO. For each image, in order to compare fairly with previous work, we extract $M = 30$ objects with high confidence scores (which is the same with [22]). To train the language model, we use the Adam optimizer [73] with a base learning rate of 5e-4. The momentum and weight decay are 0.8 and 0.999 respectively. We train the model up to 50 epochs with early stopping. We first train the language LSTM with conventional cross-entropy loss and then retrain the model with sentence-level reward loss. For a fair comparison with previous work [25], [45], we utilize the CIDEr score to optimize our model.

IV. Experiments

A. Datasets and Evaluation Metrics

We conduct experiments on the Microsoft COCO dataset [72] and the VG-COCO dataset (which is the intersection of the Visual Genome (VG) dataset [59] and the Microsoft COCO dataset) to evaluate the performance of our method.

Microsoft COCO: The Microsoft COCO dataset is currently the most commonly used and the largest corpus for image captioning. This dataset contains 82,783 images for training, 40,504 images for validation and 40,775 images for testing, where each image is associated with 5 sentences. In order to conveniently compare with previous methods, we use the same splits as [40]. It contains 82,783 images from the training set for training, 5,000 images from the validation set for validation, as well as another 5,000 images from the validation set for testing.

VG-COCO: The VG dataset has a total of 108,077 images with full annotation information such as objects, attributes, and relationships among objects [59]. In order to obtain a dataset with both captions and relationships for images, the intersection of VG and MS COCO is adopted in our paper, which is named as VG-COCO. It contains a total of 51,208 images. In our paper, we use the 33,848 images which are contained in the training set of MS COCO for training, and we select 2,000 images for validation and 2,000 for testing from the remaining images.

For evaluation, we use the most commonly used metrics: BLEU [74], METEOR [75], ROUGE-L [76], CIDEr [77], SPICE [78] and WMD [79]. BLEU evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references, and METEOR evaluates a generated sentence by computing a score based on word-level matches between the generation and a set of references. ROUGE-L counts the number of overlapping units between the generated description and its references. CIDEr uses human consensus to evaluate the similarity of a generated sentence against the references. SPICE computes caption similarity based on the agreement of the scene graph tuples of the candidate sentence and all reference sentences. WMD is calculated based on word2vec embeddings of the words. Similar to the work of [79], we also convert the distance scores to similarities by using a negative exponential.

B. Relationships for Image Captioning

Scene graphs not only depict object entities in images but also present pairwise relationships. They are structured representations which contain rich relationships, providing abundant information for describing images. In this subsection, we implement experiments on the VG-COCO dataset to verify that relationships are useful for image captioning.

We first extract N relationships with high scores for each image I, and then we decode them into relationship feature vectors $R_I = [r_1,\ r_2,\ ...\ r_N]$ with the method introduced in Section III-B. Based on these features, we use semantic attention (Att-S, as introduced in Section III-C) to generate captions. Because each image has an average of 6.2 ground truth relationships [30], we set the numbers of relationships as 5, 10, and 15. As VG-COCO is proposed for the first time in our paper, we use the method proposed by Vinyals [20] as our baseline (Google NIC). Specifically, given an image, we first extract CNN features of the last fully connected layer and then feed these features into LSTM to generate captions. Fig. 6 illustrates the performance with various numbers of relationships as well as the baseline on the test set of VG-COCO. It can be observed that representing the image with semantic relationship features is considerably better than the CNN features of the whole image. By automatically attending to relationships which describe the interactions of local regions, our method can obtain better performance. The results demonstrate that relationships are useful for image captioning. With the increase of the relationships, more details of the image are revealed, and thus, the performance increases. However, excessive relationships will bring noise, leading to poorer performance. As shown in Fig. 6, when the number of relationships is 10, the model achieves the best performance. Therefore, we set $N = 10$ in the following subsections.

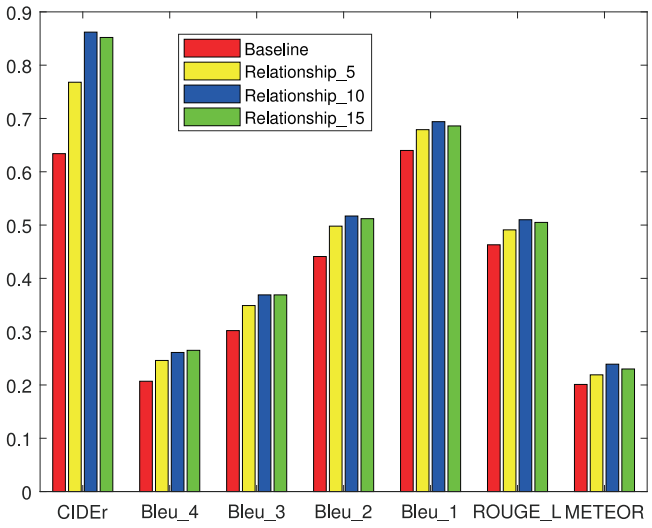

Fig. 6. The effect of the number of relationships for generating natural language descriptions on the VG-COCO dataset. Baseline denotes the method which uses the CNN features of the whole image to generate captions. Relationship_N denotes our methods which use top-N relationships to generate captions.

As described in Section III-B, after obtaining the feature vector for each word, we use mean pooling to obtain the integral one to represent each relationship triple (i.e., mean pooling). We also try other approaches to encode each relationship triple. First, we intuitively use max pooling to fuse these three vectors (i.e., max pooling). Second, we feed the relationship triple into an LSTM and use the last hidden state of the LSTM to represent it (i.e., LSTM). The results are shown in Table I. Because our relationship triples are fixe-length sequences, the mean-pooling-based method has the best performance, which is slightly superior to the LSTM-based encoding approach. In the following subsections, we use mean pooling to encode the relationship triples.

C. Effectiveness of the Proposed Method

In this subsection, we evaluate the effectiveness of our method which utilizes hierarchical-attention-based method to leverage visual features and semantic relationship features from scene graphs for image captioning. The results on the test set of VG-COCO are shown in Table II.

We report the approach which only uses visual attention (Att-V) on the visual features (i.e., v_t) from the nodes of scene graphs to generate image descriptions (Obj). For comparison, we also use semantic relationship features (i.e., s_t) alone to generate captions (Rel, the same method in Section IV-B). We experiment with two approaches to combine visual features and semantic features. The first one is to directly concatenate them (Obj+Rel-C) and the second one is to employ our hierarchical-attention-based model (Obj+Rel-A). As shown in Table II, the results

TABLE I
RESULTS OF DIFFERENT ENCODING METHODS FOR OBTAINING SEMANTIC RELATIONSHIP FEATURES ON THE TEST SET OF VG-COCO DATASET. BLEU, METEOR, ROUGE-L, CIDER, SPICE, AND WMD ARE USED AS THE METRICS, WHERE B-N DENOTES BLEU SCORES WITH N-GRAM (N = 1, 2, 3, 4). HIGHER IS BETTER

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	SPICE	WMD
LSTM	0.693	0.515	0.368	0.260	0.238	0.509	0.859	0.168	0.122
Max Pooling	0.689	0.512	0.365	0.258	0.237	0.507	0.849	0.164	0.121
Mean Pooling	**0.694**	**0.517**	**0.369**	**0.261**	**0.239**	**0.510**	**0.862**	**0.169**	**0.122**

TABLE II
AUTOMATIC METRIC SCORES OF DIFFERENT METHODS ON THE TEST SET OF VG-COCO DATASET. B-N DENOTES BLEU SCORES WITH N-GRAM (N = 1, 2, 3, 4). HIGHER IS BETTER

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	SPICE	WMD
Google NIC [20]	0.640	0.441	0.302	0.207	0.201	0.463	0.634	0.134	0.106
Obj (Ours)	0.678	0.488	0.346	0.255	0.219	0.491	0.801	0.156	0.117
Rel (Ours)	0.694	0.517	0.369	0.261	0.239	0.510	0.862	0.169	0.122
Obj+Rel-C (Ours)	0.701	0.523	0.372	0.261	0.240	0.512	0.873	0.172	0.145
Obj+Rel-A (Ours)	**0.712**	**0.533**	**0.382**	**0.270**	**0.244**	**0.518**	**0.899**	**0.175**	**0.149**
NBT† [44]	0.738	0.571	0.428	0.326	0.258	0.530	0.968	0.186	0.164
Obj-R+Rel-A (Ours)	**0.752**	**0.583**	**0.451**	**0.332**	**0.262**	**0.551**	**1.054**	**0.194**	**0.176**
Obj-R+Rel-A+CIDEr (Ours)	0.779	0.615	0.463	0.340	0.267	0.558	1.133	0.201	0.182

†indicates uses ResNet-101 features.

on the VG-COCO dataset demonstrate that our method which uses hierarchical attention based approach to leverage both visual features and semantic relationship features (Obj+Rel-A) achieves the best performance. Many conclusions can be obtained from Table II. First, as Obj and Rel are both better than Google NIC, it demonstrates that the extracted visual features and semantic relationship features provide informative cues for generating natural language descriptions. Second, because relationships inherently bundle two objects together by reasoning their interactions, they inherently contain semantic information and provide an in-depth understanding of images. For these reasons, as the results show, Rel performs better than Obj. Third, because both Obj+Rel-A and Obj+Rel-C perform better than individual Obj and Rel, it is demonstrated that visual object features (Obj) and semantic relationship features (Rel) are complementary. At last, as Obj+Rel-A is better than Obj+Rel-C, it demonstrates that directly connecting visual and semantic information is less informative than using attention mechanism. For words such as "woman" and "computer", the language model requires more information from the visual objects. However, for words such as "sitting" and "down", the language model requires more information from the semantic relationship features. Hierarchical attention can adaptively attend to these two kinds of features, thus fusing useful features, so Obj+Rel-A performs better than Obj+Rel-C. The results demonstrate the effectiveness of our hierarchical-attention-based approach. With complementary visual object features and semantic relationship features from scene graphs, our method can generate more accurate captions.

We also compare our method with NBT [44], which utilizes objects in images to generate natural language explicitly grounded in entities found by object detectors. As shown in Table II, Obj+Rel-A obtains slightly worse performance than NBT. The reasons come from two sides. First, NBT utilizes a two-stage process to generate language. Compared with our method, NBT contains an additional module which decides whether to generate a word from the textual vocabulary or generate a visual slot at each time step. Second, we use visual features extracted from VGG-16 [14], while NBT uses ResNet-101. For a fair comparison with NBT [44], we use ResNet-101 to extract visual features for the detected objects. With better visual features, Obj-R+Rel-A surpasses NBT with a gain of 0.014 in the BLEU-1 metric. Furthermore, we also utilize the CIDEr score to optimize our model. After optimizing the CIDER score, Obj-R+Rel-A+CIDEr obtains better performance. For example, compared with NBT, we obtain a gain of 0.041 in the BLEU-1 metric and a gain of 0.165 in the CIDEr metric.

D. Comparison With State-of-the-Art

We compare the proposed method which use both visual and semantic relationship features from scene graphs for image captioning with state-of-the-art image captioning models on the MS COCO dataset. The results are shown in Table III. It can be observed that the proposed method achieves promising results. This is due to the fact that our model exploits visual features as well as complementary semantic relationship features from scene graphs. The method proposed by Jin *et al.* [22] (i.e., ObjOnly) is similar to our Obj setting, which uses visual features extracted from objects for image captioning. When using visual object features alone (Obj), we obtain almost the same performance as ObjOnly [22]. After incorporating semantic relationship features, our Obj+Rel-A model surpasses ObjOnly with large gains (e.g., 0.127 in CIDEr, 0.057 in BLEU-2). Meanwhile, our method also outperforms the approach proposed by You *et al.* [23] which uses semantic concepts (Sem_ATT) with large gains (e.g., 0.013 in METEOR, 0.39 in BLEU-2). With the same visual features (i.e., ResNet-101 features), Obj-R+Rel-A surpasses Adaptive-Attn [37] with a gain of 0.025 in the BLEU-1 metric and a gain of 0.018 in the CIDEr metric. Meanwhile, it surpasses NBT [44] with a gain of 0.012 in the BLEU-1 metric. After optimized for the CIDEr score, our method (i.e.,

TABLE III
RESULTS COMPARED WITH STATE-OF-THE-ART METHODS ON THE TEST PORTION OF KARPATHY'S SPLITS ON MS COCO. B-N DENOTES BLEU SCORES WITH N-GRAM (N = 1, 2, 3, 4). "-" INDICATES THE RESULTS ARE NOT REPORTED IN THE CORRESPONDING REFERENCES. HIGHER IS BETTER

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	SPICE	WMD
Google NIC [20]	-	-	-	0.277	0.237	-	-	-	-
LRCN [38]	0.628	0.442	0.304	0.210	-	-	-	-	-
m-RNN [41]	0.670	0.490	0.350	0.250	-	-	-	-	-
ObjOnly [22]	0.697	0.519	0.381	0.282	0.235	0.509	0.838	-	-
CNN+Attn [43]	0.722	0.553	0.418	0.316	0.250	0.531	0.952	-	-
SCA-CNN [80]	0.705	0.533	0.397	0.298	0.242	-	-	-	-
SAT-Soft [21]	0.707	0.492	0.344	0.243	0.239	-	-	-	-
SAT-Hard [21]	0.718	0.504	0.357	0.250	0.230	-	-	-	-
Sem_ATT [23]	0.709	0.537	0.402	0.304	0.243	-	-	-	-
Adap [3]	0.728	0.527	0.376	0.268	0.238	-	-	-	-
Obj (Ours)	0.701	0.524	0.386	0.284	0.235	0.512	0.836	0.174	0.148
Rel (Ours)	0.718	0.536	0.420	0.302	0.240	0.522	0.926	0.186	0.153
Obj+Rel-A (Ours)	**0.732**	**0.576**	**0.432**	**0.328**	**0.256**	**0.534**	**0.965**	**0.192**	**0.162**
Attributes†[28]	0.735	0.567	0.430	0.326	0.255	0.540	1.002	-	-
SCN† [24]	0.728	0.566	0.433	0.330	0.257	-	1.012	-	-
Adaptive-Attn†[37]	0.742	0.580	0.439	0.332	0.266	-	1.085	0.195	-
NBT†[44]	0.755	-	-	**0.347**	**0.271**	-	1.072	**0.201**	-
Obj-R+Rel-A (Ours)	**0.767**	**0.598**	**0.453**	0.338	0.262	**0.549**	**1.103**	0.198	**0.180**
Self-Crit* [45]	-	-	-	0.342	0.267	0.557	1.140	-	-
PG-BCMR* [46]	0.754	0.591	0.445	0.332	0.257	0.550	1.013	-	-
Up-Down* [25]	**0.798**	-	-	0.363	**0.277**	**0.569**	1.201	0.214	-
Obj-R+Rel-A+CIDEr (Ours)	0.792	**0.632**	**0.483**	**0.363**	0.276	0.568	**1.202**	**0.214**	**0.186**

†indicates uses ResNet-101 features and*indicates optimizing the CIDEr score.

TABLE IV
LEADERBOARD OF THE PUBLISHED IMAGE CAPTIONING MODELS ON THE ONLINE MS COCO TESTING SERVER. B-N DENOTES BLEU SCORES WITH N-GRAM (N = 1, 2, 3, 4). OUR METHOD WHICH UTILIZES VISUAL OBJECT FEATURES AND SEMANTIC RELATIONSHIP FEATURES ACHIEVES COMPARABLE PERFORMANCE. HIGHER IS BETTER

Method	B-1		B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
CNN+Attn [43]	0.708	0.883	0.534	0.786	0.389	0.667	0.280	0.545	0.241	0.321	0.517	0.657	0.872	0.893
m-RNN [41]	0.716	0.890	0.545	0.798	0.404	0.687	0.299	0.575	0.242	0.325	0.521	0.666	0.917	0.935
LRCN [38]	0.718	0.895	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.528	0.678	0.921	0.934
SCA-CNN [80]	0.712	0.894	0.542	0.802	0.404	0.691	0.302	0.579	0.244	0.331	0.524	0.674	0.912	0.921
SAT_Hard [21]	0.705	0.881	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.516	0.654	0.865	0.893
Adaptive-Attn [37]	0.748	0.920	0.584	0.845	0.444	0.744	0.336	0.637	0.264	0.359	0.550	0.705	1.042	1.059
PG-BCMR [46]	0.754	-	0.591	-	0.445	-	0.332	-	0.257	-	0.550	-	1.103	-
Self-Crit [45]	0.781	0.937	0.619	0.860	0.470	0.759	0.352	0.645	0.270	0.355	0.563	0.707	1.147	1.167
Obj-R+Rel-A+CIDEr (Ours)	**0.792**	**0.944**	**0.626**	**0.872**	**0.475**	**0.771**	**0.354**	**0.658**	**0.273**	**0.361**	**0.562**	**0.712**	**1.151**	**1.173**
Google NIC†[20]	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946
SAT† [21]	0.731	0.901	0.565	0.816	0.424	0.710	0.316	0.600	0.251	0.336	0.535	0.683	0.944	0.959
Up-Down† [25]	**0.802**	**0.952**	**0.641**	**0.888**	**0.491**	**0.794**	**0.369**	**0.685**	**0.276**	**0.367**	**0.571**	**0.724**	**1.179**	**1.205**

†indicates that the results are obtained from an ensemble model.

Obj-R+Rel-A+CIDEr) obtains better performance. Obj-R+Rel-A+CIDEr surpasses PG-BCMR [46] with a gain of 0.038 in the BLEU-1 metric and a gain of 0.089 in the CIDEr metric. What is more, Obj-R+Rel-A+CIDEr achieves comparable performance with Up-Down [24], and it even achieves better performance in the CIDEr metric.

We also evaluate our model on the MS COCO Image Challenge set by uploading results to the official test sever. The results are shown in Table IV, where c5 and c40 indicate the numbers of reference captions for evaluation. It can be observed that our approach achieves very competitive performance, compared to the state-of-the-art approaches. Our model outperforms the other models in most cases. This is due to the fact that our method exploits visual object features, semantic relationship features, and hierarchical attention. The reason why our method is slightly worse than Up-Down [25] is that Up-Down utilizes an ensemble of 4 models, whereas our Obj-R+Rel-A+CIDEr is a single one. Ensemble models can always obtain better results than single model. The results on the online MS COCO test server demonstrate the effectiveness of our proposed method.

E. Qualitative Results

To demonstrate that our method which utilizes structured scene graphs can generate better image descriptions, some qualitative results are presented in Fig. 7. By integrating both visual features and semantic relationship features, our model can generate more comprehensive captions. For example, as shown in the bottom of the first column of Fig. 7, even though Obj which exploits visual features from the image can predict the correct object categories (i.e., *man* and *dog*), it is unable to describe them properly. The expression of *"man with dog"* is somewhat improper. However, with both features from objects and their

Fig. 7. Examples which are generated by different methods in the test set of the VG-COCO dataset. The scores in the parentheses are the METEOR values for the generated captions, which are evaluated over the corresponding the ground-truth (GT) captions.

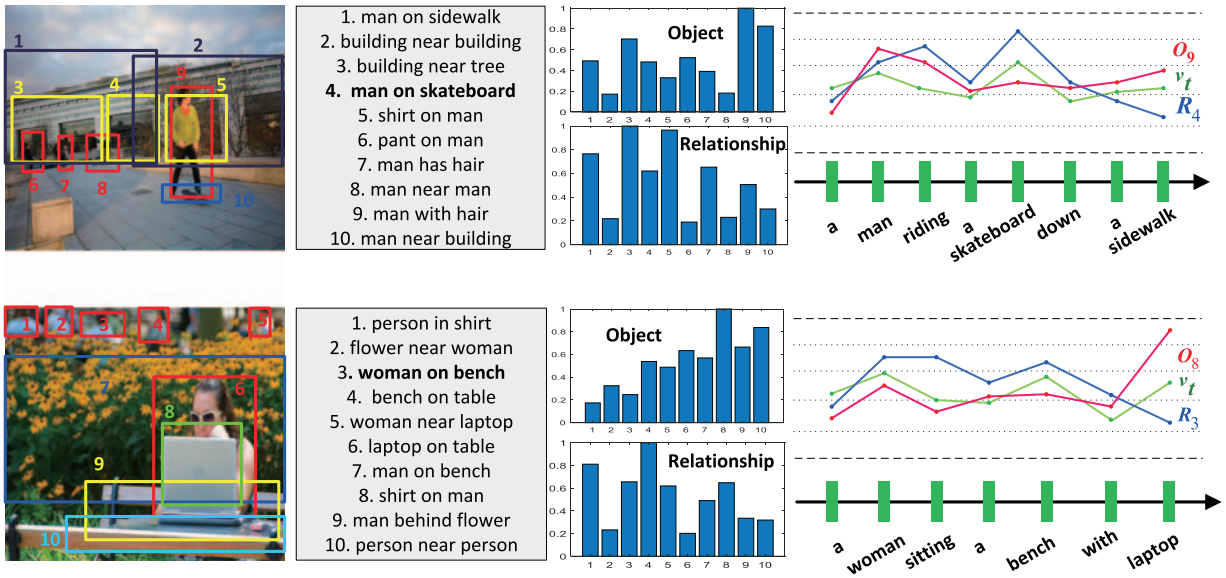

Fig. 8. Examples of generated captions and attention weights. The first and second columns show the original images and the extracted objects with the corresponding relationships. The third column shows the weight for each object and relationship (top10 with high relevance scores) when generating the whole target caption. We first accumulate the relevance scores for each item (i.e., object scores from visual attention and relationship scores from semantic attention) at each time step, and then min-max normalization of the accumulated scores is used to represent the final weights (Indexes are just for convenient display). The fourth column shows the changes of the attention weights for some objects (i.e., O_i denotes the object marked as i), relationships (i.e., R_i denotes the relationship marked as i), and the weighted visual context v_t at each time step (Best viewed in color).

relationships, Obj+Rel-A not only predicts the correct object categories, but also describes them accurately (*man holding dog*). The other examples in Fig. 7 also demonstrate that adaptively attending to visual object features and semantic relationship features from structured scene graphs can generate more accurate image captions.

F. Attention Analysis

In this subsection, we quantitatively show that our hierarchical-attention-based feature fusion module can automatically learn important cues from scene graphs for describing images. As many approaches [3], [21], [23], [37] have shown

that an attention-based encoder can selectively attend to visual regions or semantic concepts, we illustrate that our model can assign different weights for different relationships and objects, which pinpoint different aspects of images. We first accumulate the relevance scores for each relationship from the semantic attention at each time step during the caption generation procedure, and then min-max normalization of the accumulated scores is used to represent the final weights. In the same spirit, we can obtain the weight for each object from the visual attention. The results show that our model can adaptively attend to important information from scene graphs. For example, as shown in the first row of Fig. 8, when generating the caption "a man riding a skateboard down a sidewalk", the relationship between the man and the skateboard (i.e., relationship 4) is selected as the key formation as it has the biggest weight. Meanwhile, corresponding visual objects (i.e., object 9 and 10) are also selected with large weights. The fourth column in Fig. 8 shows the changes in the attention weights for some objects, relationships, and the weighted visual context v_t at each step of the language generation process. Several conclusions can be drawn. First, both Att-V and Att-S can selectively attend to relevant cues. For example, as shown in the second row and the fourth column of Fig. 8, to generate the word "laptop", Att-V mainly focuses on the object instance of "laptop". For the word "sitting", Att-S mainly focuses on the relationship of "woman on bench". Second, Att-M can adaptively fuse the visual v_t and semantic s_t context. For example, for the visual words (i.e., "woman", "bench", and "laptop"), Att-M mainly focuses on visual context v_t. Conversely, for words such as "sitting" and "with", Att-M mainly focuses on semantic context because the attention weights for v_t are small. By assigning different weights for different objects, relationships, and the weighted context, our hierarchical-attention-based model can automatically learn useful information from structured scene graphs for image captioning.

V. Conclusion

In this work, we propose a framework that is based on scene graphs for image captioning. Scene graphs contain abundant structured information because they not only depict object entities in images but also present pairwise relationships. To leverage both visual features and semantic knowledge in structured scene graphs, we extract CNN features from the bounding box offsets of object entities for visual representations, and we extract semantic relationship features from triples for semantic representations. After obtaining these features, we introduce a hierarchical-attention-based module to learn discriminative features for word generation at each time step. Our model is able to learn important cues from scene graphs and achieves promising results. In future work, we will learn scene graphs for images with the supervision of the target captions to obtain better image representations for generating natural language descriptions.

References

[1] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[2] L. Li *et al.*, "GLA: Global and local attention for image description," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 726–737, Mar. 2018.

[3] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H. T. Shen, "Adaptively attending to visual attributes and linguistic knowledge for captioning," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1345–1353.

[4] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1549–1557.

[5] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu, "StructCap: Structured semantic embedding for image captioning," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 46–54.

[6] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional LSTMs," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 988–997.

[7] X. Li and S. Jiang, "Bundled object context for referring expressions," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2749–2760, Oct. 2018.

[8] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[9] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[10] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, " Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.

[11] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3674–3683.

[12] D. Fried *et al.*, "Speaker-follower models for vision-and-language navigation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3318–3329.

[13] X. Wang, W. Xiong, H. Wang, and W. Y. Wang, "Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 37–53.

[14] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

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

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

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

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[21] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[22] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image caption with region-based attention and scene factorization," 2015, arXiv:1506.06272. [Online]. Available: http://arxiv.org/abs/1506.06272

[23] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.

[24] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1141–1150.

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

[26] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[27] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 203–212.

[28] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4904–4912.

[29] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2407–2415.

[30] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3097–3106.

[31] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1270–1279.

[32] J. Johnson *et al.*, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3668–3678.

[33] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Generalizing image captions for image-text parallel corpus," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistic*, 2013, pp. 790–796.

[34] V. Ordonez *et al.*,"Large scale retrieval and generation of image descriptions," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 46–59, 2016.

[35] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.

[36] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[37] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3242–3250.

[38] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[39] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2989–2998.

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

[41] J. Mao *et al.*, "Deep captioning with multimodal recurrent neural networks M-RNN," *Proc. Int. Conf. Learn. Represent.*, 2015.

[42] X. Li, X. Song, L. Herranz, Y. Zhu, and S. Jiang, "Image captioning with both object and scene information," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 1107–1110.

[43] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5561–5570.

[44] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7219–7228.

[45] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1179–1195.

[46] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proc. IEEE Int. Conf. Comp. Vis*, 2017, pp. 873–881.

[47] K. Lin, D. Li, X. He, Z. Zhang, and M.-T. Sun, "Adversarial ranking for language generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3155–3165.

[48] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 158–172.

[49] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3-D geometric phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 33–40.

[50] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.

[51] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4408–4417.

[52] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1068–1076.

[53] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1946–1955.

[54] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: Visual phrase guided convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7244–7253.

[55] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3298–3308.

[56] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5532–5540.

[57] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5831–5840.

[58] Y.-S. Wang, C. Liu, X. Zeng, and A. Yuille, "Scene graph parsing as dependency parsing," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technologies*, 2018, pp. 397–407.

[59] R. Krishna *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[60] M. Klawonn and E. Heim, "Generating triples with adversarial networks for scene graph construction," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6992–6999.

[61] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3233–3241.

[62] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1219–1228.

[63] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[64] C. Liu, J. Mao, F. Sha, and A. L. Yuille, "Attention correctness in neural image captioning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4176–4182.

[65] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2361–2369.

[66] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention LSTM networks for video captioning," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 537–545.

[67] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385. Berlin, Germany: Springer, 2012.

[68] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. SSST-8, 8th Workshop Syntax*, Semantics and Structure in Statistical Translation, 2014, pp. 103–111.

[69] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[70] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5005–5013.

[71] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

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

[73] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[74] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistic*, 2002, pp. 311–318.

[75] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Statistical Mach. Translation Assoc. Comput. Linguistic*, 2007, pp. 228–231.

[76] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop, Text Summarization Branches Out*, 2004, pp. 74–81.

[77] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[78] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.

[79] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 199–209.

[80] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6298–6306.

Xiangyang Li received the B.E. degree from the Wuhan Institute of Technology, Wuhan, China, in 2012, the M.E. degree from the Capital Normal University, Beijing, China, in 2015. He is currently working toward the Ph.D. degree with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research interests include large-scale image classification, joint learning with language and vision, computer vision, and pattern recognition.

Shuqiang Jiang (SM'08) is a Professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing, China, and a Professor with the University of CAS, Beijing, China. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and intelligent understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, National Natural Science Foundation of China Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the China Computer Federation Award of Science and Technology in 2012.

Prof. Jiang is the senior member of CCF, the member of ACM, an Associate Editor for ACM Transactions on Multimedia Computing, Communications, and Applications, the IEEE Multimedia, Multimedia Tools and Applications. He is the Vice Chair of the ICircuits and Systems Society Beijing Chapter, the Vice Chair of ACM SIGMM China chapter. He has served as an organization member of more than 20 academic conferences, including the General Chair of the International Conference on Internet Multimedia Computing and Service (ICIMCS 2015), a Program Chair of ICIMCS 2010, Grand Challenges Chair of ACM Mutlimedia 2018, Special Session Chair of ACM International Conference on Multimedia Retrieval 2018, and PCM 2008, etc. He has also served as a TPC member for many conferences, including ACM Multimedia, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, International Joint Conference on Artifical intelligence, IEEE International Conference on Multimedia and Expo, International Conference on Image Processing, etc.