Exploring Visual Relationship for Image Captioning

Ting Yao[1(✉)], Yingwei Pan[1], Yehao Li[2], and Tao Mei[1]

[1] JD AI Research, Beijing, China
tingyao.ustc@gmail.com, panyw.ustc@gmail.com, tmei@live.com
[2] Sun Yat-sen University, Guangzhou, China
yehaoli.sysu@gmail.com

Abstract. It is always well believed that modeling relationships between objects would be helpful for representing and eventually describing an image. Nevertheless, there has not been evidence in support of the idea on image description generation. In this paper, we introduce a new design to explore the connections between objects for image captioning under the umbrella of attention-based encoder-decoder framework. Specifically, we present Graph Convolutional Networks plus Long Short-Term Memory (dubbed as GCN-LSTM) architecture that novelly integrates both semantic and spatial object relationships into image encoder. Technically, we build graphs over the detected objects in an image based on their spatial and semantic connections. The representations of each region proposed on objects are then refined by leveraging graph structure through GCN. With the learnt region-level features, our GCN-LSTM capitalizes on LSTM-based captioning framework with attention mechanism for sentence generation. Extensive experiments are conducted on COCO image captioning dataset, and superior results are reported when comparing to state-of-the-art approaches. More remarkably, GCN-LSTM increases CIDEr-D performance from 120.1% to 128.7% on COCO testing set.

Keywords: Image captioning · Graph convolutional networks
Visual relationship · Long short-term memory

1 Introduction

The recent advances in deep neural networks have convincingly demonstrated high capability in learning vision models particularly for recognition. The achievements make a further step towards the ultimate goal of image understanding, which is to automatically describe image content with a complete and natural sentence or referred to as image captioning problem. The typical solutions [7,34,37,39] of image captioning are inspired by machine translation and equivalent to translating an image to a text. As illustrated in Fig. 1(a) and (b), a Convolutional Neural Network (CNN) or Region-based CNN (R-CNN) is usually

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11218, pp. 711–727, 2018.
https://doi.org/10.1007/978-3-030-01264-9_42

exploited to encode an image and a decoder of Recurrent Neural Network (RNN) w/ or w/o attention mechanism is utilized to generate the sentence, one word at each time step. Regardless of these different versions of CNN plus RNN image captioning framework, a common issue not fully studied is how visual relationships should be leveraged in view that the mutual correlations or interactions between objects are the natural basis for describing an image.

Visual relationships characterize the interactions or relative positions between objects detected in an image. The detection of visual relationships involves not only localizing and recognizing objects, but also classifying the interaction (predicate) between each pair of objects. In general, the relationship can be represented as ⟨*subject-predicate-object*⟩, e.g., ⟨*man-eating-sandwich*⟩ or ⟨*dog-inside-car*⟩. In the literature, it is well recognized that reasoning such visual relationships is crucial to a richer semantic understanding [19,23] of the visual world. Nevertheless, the fact that the objects could be with a wide range of scales, at arbitrary positions in an image and from different categories results in difficulty in determining the type of relationships. In this paper, we take the advantages of the inherent relationships between objects for interpreting the images holistically and novelly explore the use of visual connections to enhance image encoder for image captioning. Our basic design is to model the relationships on both semantic and spatial levels, and integrate the connections into image encoder to produce relation-aware region-level representations. As a result, we endow image representations with more power when feeding into sentence decoder.

By consolidating the idea of modeling visual relationship for image captioning, we present a novel Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) architecture, as conceptually shown in Fig. 1(c). Specifically, Faster R-CNN is firstly implemented to propose a set of salient image regions. We build semantic graph with directed edges on the detected regions, where the vertex represents each region and the edge denotes the relationship (predicate) between each pair of regions which is predicted by semantic relationship detector learnt on Visual Genome [16]. Similarly, spatial graph is also constructed on the regions and the edge between regions models relative geometrical relationship. Graph Convolutional Networks are then exploited to enrich region representations with visual relationship in the structured semantic and spatial graph respectively. After that, the learnt relation-aware region represen-

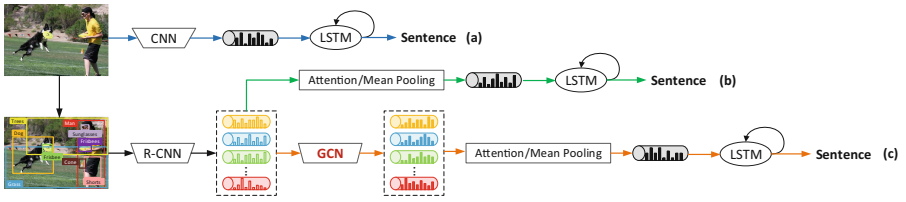

Fig. 1. Visual representations generated by image encoder in (a) CNN plus LSTM, (b) R-CNN plus LSTM, and (c) our GCN-LSTM for image captioning.

tations on each kind of relationships are feed into one individual attention LSTM decoder to generate the sentence. In the inference stage, to fuse the outputs of two decoders, we linearly average the predicted score distributions on words from two decoders at each time step and pop out the word with the highest probability as the input word to both decoders at the next step.

The main contribution of this work is the proposal of the use of visual relationship for enriching region-level representations and eventually enhancing image captioning. This also leads to the elegant views of what kind of visual relationships could be built between objects, and how to nicely leverage such visual relationships to learn more informative and relation-aware region representations for image captioning, which are problems not yet fully understood.

2 Related Work

Image Captioning. With the prevalence of deep learning [17] in computer vision, the dominant paradigm in modern image captioning is sequence learning methods [7,34,37–40] which utilize CNN plus RNN model to generate novel sentences with flexible syntactical structures. For instance, Vinyals *et al.* propose an end-to-end neural networks architecture by utilizing LSTM to generate sentence for an image in [34], which is further incorporated with soft/hard attention mechanism in [37] to automatically focus on salient objects when generating corresponding words. Instead of activating visual attention over image for every generated word, [24] develops an adaptive attention encoder-decoder model for automatically deciding when to rely on visual signals/language model. Recently, in [35,39], semantic attributes are shown to clearly boost image captioning when injected into CNN plus RNN model and such attributes can be further leveraged as semantic attention [40] to enhance image captioning. Most recently, a novel attention based encoder-decoder model [2] is proposed to detect a set of salient image regions via bottom-up attention mechanism and then attend to the salient regions with top-down attention mechanism for sentence generation.

Visual Relationship Detection. Research on visual relationship detection has attracted increasing attention. Some early works [9,10] attempt to learn four spatial relations (i.e., "above", "below", "inside" and "around") to improve segmentation. Later on, semantic relations (e.g., actions or interactions) between objects are explored in [6,32] where each possible combination of semantic relation is taken as a visual phrase class and the visual relationship detection is formulated as a classification task. Recently, quite a few works [5,19,23,29,36] design deep learning based architectures for visual relationship detection. [36] treats visual relationship as the directed edges to connect two object nodes in the scene graph and the relationships are inferred along the processing of constructing scene graph in an iterative way. [5,19] directly learn the visual features for relationship prediction based on additional union bounding boxes which cover object and subject together. In [23,29], the linguistic cues of the participating objects/captions are further considered for visual relationship detection.

Summary. In short, our approach in this paper belongs to sequence learning method for image captioning. Similar to previous approaches [2,8], GCN-LSTM explores visual attention over the detected image regions of objects for sentence generation. The novelty is on the exploitation of semantic and spatial relations between objects for image captioning, that has not been previously explored. In particular, both of the two kinds of visual relationships are seamlessly integrated into LSTM-based captioning framework via GCN, targeting for producing relation-aware region representations and thus potentially enhancing the quality of generated sentence through emphasizing the object relations.

3 Image Captioning by Exploring Visual Relationship

We devise our Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) architecture to generate image descriptions by additionally incorporating both semantic and spatial object relationships. GCN-LSTM firstly utilizes an object detection module (e.g., Faster R-CNN [30]) to detect objects within images, aiming for encoding and generalizing the whole image into a set of salient image regions containing objects. Semantic and spatial relation graphs are then constructed over all the detected image regions of objects based on their semantic and spatial connections, respectively. Next, the training of GCN-LSTM is performed by contextually encoding the whole image region set with semantic or spatial graph structure via GCN, resulting in relation-aware region representations. All of encoded relation-aware region representations are further injected into LSTM-based captioning framework, enabling region-level attention mechanism for sentence generation. An overview of our image captioning architecture is illustrated in Fig. 2.

3.1 Problem Formulation

Suppose we have an image I to be described by a textual sentence $\mathcal{S}$, where $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$ consisting of N_s words. Let $\mathbf{w}_t \in \mathbb{R}^{D_s}$ denote the D_s-dimensional textual feature of the t-th word in sentence $\mathcal{S}$. Faster R-CNN is firstly leveraged to produce the set of detected objects $\mathcal{V} = \{v_i\}_{i=1}^{K}$ with K image regions of objects in I and $\mathbf{v}_i \in \mathbb{R}^{D_v}$ denotes the D_v-dimensional feature of each image region. Furthermore, by treating each image region v_i as one vertex, we can construct semantic graph $\mathcal{G}_{sem} = (\mathcal{V}, \mathcal{E}_{sem})$ and spatial graph $\mathcal{G}_{spa} = (\mathcal{V}, \mathcal{E}_{spa})$, where $\mathcal{E}_{sem}$ and $\mathcal{E}_{spa}$ denotes the set of semantic and spatial relation edges between region vertices, respectively. More details about how we mine the visual relationships between objects and construct the semantic and spatial graphs will be elaborated in Sect. 3.2.

Inspired by the recent successes of sequence models leveraged in image/video captioning [26,27,34] and region-level attention mechanism [2,8], we aim to formulate our image captioning model in a R-CNN plus RNN scheme. Our R-CNN plus RNN method firstly interprets the given image as a set of image regions with R-CNN, then uniquely encodes them into relation-aware features conditioned on

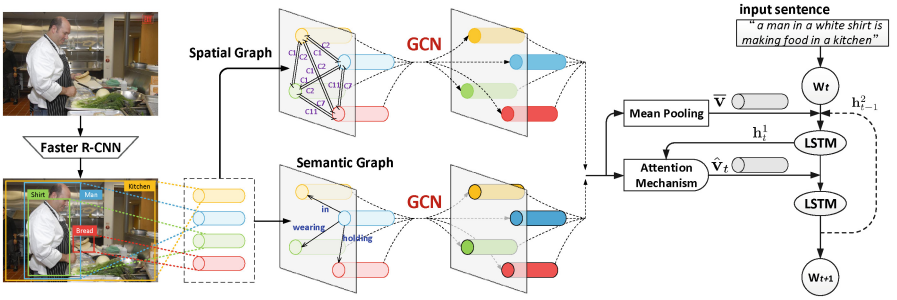

Fig. 2. An overview of our Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) for image captioning (better viewed in color). Faster R-CNN is first leveraged to detect a set of salient image regions. Next, semantic/spatial graph is built with directional edges on the detected regions, where the vertex represents each region and the edge denotes the semantic/spatial relationship in between. Graph Convolutional Networks (GCN) is then exploited to contextually encode regions with visual relationship in the structured semantic/spatial graph. After that, the learnt relation-aware region-level features from each kind of graph are feed into one individual attention LSTM decoder for sentence generation. In the inference stage, we adopt a late fusion scheme to linearly fuse the results from two decoders.

semantic/spatial graph, and finally decodes them to each target output word via attention LSTM decoder. Derived from the idea of Graph Convolutional Networks [15,25], we leverage a GCN module in image encoder to contextually refine the representation of each image region, which is endowed with the inherent visual relationships between objects. Hence, the sentence generation problem we explore here can be formulated by minimizing the following energy loss function:

$$E(\mathcal{V}, \mathcal{G}, \mathcal{S}) = -\log \Pr\left(\mathcal{S}|\mathcal{V}, \mathcal{G}\right), \tag{1}$$

which is the negative log probability of the correct textual sentence given the detected image regions of objects $\mathcal{V}$ and constructed relation graph $\mathcal{G}$. Note that we use $\mathcal{G} \in \{\mathcal{G}_{sem}, \mathcal{G}_{spa}\}$ for simplicity, i.e., $\mathcal{G}$ denotes either semantic graph $\mathcal{G}_{sem}$ or spatial graph $\mathcal{G}_{spa}$. Here the negative log probability is typically measured with cross entropy loss, which inevitably results in the discrepancy of evaluation between training and inference. Accordingly, to further boost our captioning model by amending such discrepancy, we can directly optimize the LSTM with expected sentence-level reward loss as in [18,22,31].

3.2 Visual Relationship Between Objects in Images

Semantic Object Relationship. We draw inspiration from recent advances in deep learning based visual relationship detection [5,19] and simplify it as a classification task to learn semantic relation classifier on visual relationship benchmarks (e.g., Visual Genome [16]). The general expression of semantic relation is

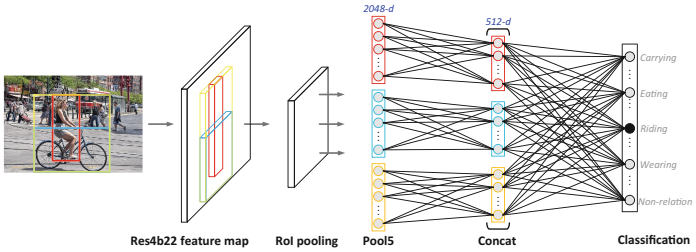

Fig. 3. Detection model for semantic relation ⟨*subject-predicate-object*⟩ (red: region of subject noun, blue: region of object noun, yellow: the union bounding box). (Color figure online)

⟨*subject-predicate-object*⟩ between pairs of objects. Note that the semantic relation is directional, i.e., it relates one object (subject noun) and another object (object noun) via a predicate which can be an action or interaction between objects. Hence, given two detected regions of objects v_i (subject noun) and v_j (object noun) within an image I, we devise a simple deep classification model to predict the semantic relation between v_i and v_j depending on the union bounding box which covers the two objects together.

Figure 3 depicts the framework of our designed semantic relation detection model. In particular, the input two region-level features $\mathbf{v}_i$ and $\mathbf{v}_j$ are first separately transformed via an embedding layer, which are further concatenated with the transferred region-level feature $\mathbf{v}_{ij}$ of the union bounding box containing both v_i and v_j. The combined features are finally injected into the classification layer that produces softmax probability over N_{sem} semantic relation classes plus a non-relation class, which is essentially a multi-class logistic regression model. Here each region-level feature is taken from the D_v-dimensional ($D_v = 2,048$) output of Pool5 layer after RoI pooling from the Res4b22 feature map of Faster R-CNN in conjunction with ResNet-101 [11].

After training the visual relation classifier on visual relationship benchmark, we directly employ the learnt visual relation classifier to construct the corresponding semantic graph $\mathcal{G}_{sem} = (\mathcal{V}, \mathcal{E}_{sem})$. Specifically, we firstly group the detected K image regions of objects within image I into $K \times (K-1)$ object pairs (two identical regions will not be grouped). Next, we compute the probability distribution on all the $(N_{sem}+1)$ relation classes for each object pair with the learnt visual relation classifier. If the probability of non-relation class is less than 0.5, a directional edge from the region vertex of subject noun to the region vertex of object noun is established and the relation class with maximum probability is regarded as the label of this edge.

Spatial Object Relationship. The semantic graph only unfolds the inherent action/interaction between objects, while leaving the spatial relations between image regions unexploited. Therefore, we construct another graph, i.e., spatial graph, to fully explore the relative spatial relations between every two regions within one image. Here we generally express the directional spatial relation as

$\langle object_i\text{-}object_j \rangle$, which represents the relative geometrical position of object$_j$ against object$_i$. The edge and the corresponding class label for every two object vertices in spatial graph $\mathcal{G}_{spa} = (\mathcal{V}, \mathcal{E}_{spa})$ are built and assigned depending on their Intersection over Union (IoU), relative distance and angle. Detailed definition of spatial relations are shown in Fig. 4.

Concretely, given two regions v_i and v_j, the locations of them are denoted as (x_i, y_i) and (x_j, y_j), which are the normalized coordinates of the centroid of the bounding box on the image plane for v_i and v_j, respectively. We can thus achieve the IoU between v_i and v_j, relative distance d_{ij} ($d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$) and relative angle θ_{ij} (i.e., the argument of the vector from the centroid of v_i to that of v_j). Two kinds of special cases are firstly considered for classifying the spatial relation between v_i and v_j. If v_i completely includes v_j or v_i is fully covered by v_j, we establish an edge from v_i to v_j and set the label of spatial relation as "inside" (**class 1**) and "cover" (**class 2**), respectively. Except for the two special classes, if the IoU between v_i and v_j is larger than 0.5, we directly connect v_i to v_j with an edge, which is classified as "overlap" (**class 3**). Otherwise, when the ratio ϕ_{ij} between the relative distance d_{ij} and the diagonal length of the whole image is less than 0.5, we classify the edge between v_i and v_j solely relying on the size of relative angle θ_{ij} and the index of class is set as $\lceil \theta_{ij}/45^\circ \rceil + 3$ (**class 4-11**). When the ratio $\phi_{ij} > 0.5$ and IoU < 0.5, the spatial relation between them is tend to be weak and no edge is established in this case.

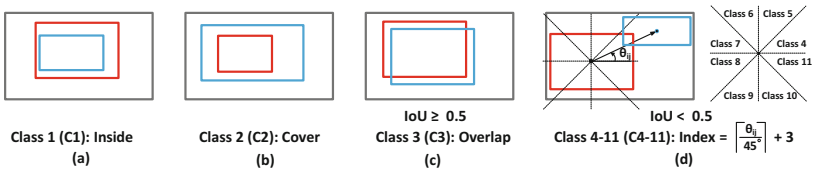

Fig. 4. Definition of eleven kinds of spatial relations $\langle object_i\text{-}object_j \rangle$ (red: region of object$_i$, blue: region of object$_j$). (Color figure online)

3.3 Image Captioning with Visual Relationship

With the constructed graphs over the detected objects based on their spatial and semantic connections, we next discuss how to integrate the learnt visual relationships into sequence learning with region-based attention mechanism for image captioning via our designed GCN-LSTM. Specifically, a GCN-based image encoder is devised to contextually encode all the image regions with semantic or spatial graph structure via GCN into relation-aware representations, which are further injected into attention LSTM for generating sentence.

GCN-based Image Encoder. Inspired from Graph Convolutional Networks for node classification [15] and semantic role labeling [25], we design a GCN-based image encoder for enriching the region-level features by capturing the

semantic/spatial relations on semantic/spatial graph, as illustrated in the middle part of Fig. 2. The original GCN is commonly operated on an undirected graph, encoding information about the neighborhood of each vertex v_i as a real-valued vector, which is computed by

$$\mathbf{v}_i^{(1)} = \rho\Big(\sum_{v_j \in \mathcal{N}(v_i)} \mathbf{W}\mathbf{v}_j + \mathbf{b}\Big), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{D_v \times D_v}$ is the transformation matrix, $\mathbf{b}$ is the bias vector and ρ denotes an activation function (e.g., ReLU). $\mathcal{N}(v_i)$ represents the set of neighbors of v_i, i.e., the region vertices have visual connections with v_i here. Note that $\mathcal{N}(v_i)$ also includes v_i itself. Although the original GCN refines each vertex by accumulating the features of its neighbors, none of the information about directionality or edge labels is included for encoding image regions. In order to enable the operation on labeled directional graph, the original GCN is upgraded by fully exploiting the directional and labeled visual connections between vertices.

Formally, consider a labeled directional graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}) \in \{\mathcal{G}_{sem}, \mathcal{G}_{spa}\}$ where $\mathcal{V}$ is the set of all the detected region vertices and $\mathcal{E}$ is a set of visual relationship edges. Separate transformation matrices and bias vectors are utilized for different directions and labels of edges, respectively, targeting for making the modified GCN sensitive to both directionality and labels. Accordingly, each vertex v_i is encoded via the modified GCN as

$$\mathbf{v}_i^{(1)} = \rho\Big(\sum_{v_j \in \mathcal{N}(v_i)} \mathbf{W}_{\mathrm{dir}(v_i,v_j)}\mathbf{v}_j + \mathbf{b}_{\mathrm{lab}(v_i,v_j)}\Big), \tag{3}$$

where $\mathrm{dir}(v_i, v_j)$ selects the transformation matrix with regard to the directionality of each edge (i.e., $\mathbf{W}_1$ for v_i-to-v_j, $\mathbf{W}_2$ for v_j-to-v_i, and $\mathbf{W}_3$ for v_i-to-v_i). $\mathrm{lab}(v_i, v_j)$ represents the label of each edge. Moreover, instead of uniformly accumulating the information from all connected vertices, an edge-wise gate unit is additionally incorporated into GCN to automatically focus on potentially important edges. Hence each vertex v_i is finally encoded via the GCN in conjunction with an edge-wise gate as

$$\begin{aligned} \mathbf{v}_i^{(1)} &= \rho\Big(\sum_{v_j \in \mathcal{N}(v_i)} g_{v_i,v_j}(\mathbf{W}_{\mathrm{dir}(v_i,v_j)}\mathbf{v}_j + \mathbf{b}_{\mathrm{lab}(v_i,v_j)})\Big), \\ g_{v_i,v_j} &= \sigma\left(\widetilde{\mathbf{W}}_{\mathrm{dir}(v_i,v_j)}\mathbf{v}_j + \widetilde{b}_{\mathrm{lab}(v_i,v_j)}\right), \end{aligned} \tag{4}$$

where g_{v_i,v_j} denotes the scale factor achieved from edge-wise gate, σ is the logistic sigmoid function, $\widetilde{\mathbf{W}}_{\mathrm{dir}(v_i,v_j)} \in \mathbb{R}^{1 \times D_v}$ is the transformation matrix and $\widetilde{b}_{\mathrm{lab}(v_i,v_j)} \in \mathbb{R}$ is the bias. Consequently, after encoding all the regions $\{\mathbf{v}_i\}_{i=1}^K$ via GCN-based image encoder as in Eq. (4), the refined region-level features $\{\mathbf{v}_i^{(1)}\}_{i=1}^K$ are endowed with the inherent visual relationships between objects.

Attention LSTM Sentence Decoder. Taking the inspiration from region-level attention mechanism in [2], we devise our attention LSTM sentence decoder

by injecting all of the relation-aware region-level features $\{\mathbf{v}_i^{(1)}\}_{i=1}^K$ into a two-layer LSTM with attention mechanism, as shown in the right part of Fig. 2. In particular, at each time step t, the attention LSTM decoder firstly collects the maximum contextual information by concatenating the input word w_t with the previous output of the second-layer LSTM unit $\mathbf{h}_{t-1}^2$ and the mean-pooled image feature $\overline{\mathbf{v}} = \frac{1}{K}\sum_{i=1}^{K}\mathbf{v}_i^{(1)}$, which will be set as the input of the first-layer LSTM unit. Hence the updating procedure for the first-layer LSTM unit is as

$$\mathbf{h}_t^1 = f_1\left(\left[\mathbf{h}_{t-1}^2, \mathbf{W}_s\mathbf{w}_t, \overline{\mathbf{v}}\right]\right), \tag{5}$$

where $\mathbf{W}_s \in \mathbb{R}^{D_s^1 \times D_s}$ is the transformation matrix for input word w_t, $\mathbf{h}_t^1 \in \mathbb{R}^{D_h}$ is the output of the first-layer LSTM unit, and f_1 is the updating function within the first-layer LSTM unit. Next, depending on the output $\mathbf{h}_t^1$ of the first-layer LSTM unit, a normalized attention distribution over all the relation-aware region-level features is generated as

$$a_{t,i} = \mathbf{W}_a\left[\tanh\left(\mathbf{W}_f\mathbf{v}_i^{(1)} + \mathbf{W}_h\mathbf{h}_t^1\right)\right], \quad \lambda_t = softmax\left(\mathbf{a}_t\right), \tag{6}$$

where $a_{t,i}$ is the i-th element of $\mathbf{a}_t$, $\mathbf{W}_a \in \mathbb{R}^{1\times D_a}$, $\mathbf{W}_f \in \mathbb{R}^{D_a \times D_v}$ and $\mathbf{W}_h \in \mathbb{R}^{D_a \times D_h}$ are transformation matrices. $\lambda_t \in \mathbb{R}^K$ denotes the normalized attention distribution and its i-th element $\lambda_{t,i}$ is the attention probability of $\mathbf{v}_i^{(1)}$. Based on the attention distribution, we calculate the attended image feature $\hat{\mathbf{v}}_t = \sum_{i=1}^{K}\lambda_{t,i}\mathbf{v}_i^{(1)}$ by aggregating all the region-level features weighted with attention. We further concatenate the attended image feature $\hat{\mathbf{v}}_t$ with $\mathbf{h}_t^1$ and feed them into the second-layer LSTM unit, whose updating procedure is thus given by

$$\mathbf{h}_t^2 = f_2\left(\left[\hat{\mathbf{v}}_t, \mathbf{h}_t^1\right]\right), \tag{7}$$

where f_2 is the updating function within the second-layer LSTM unit. The output of the second-layer LSTM unit $\mathbf{h}_t^2$ is leveraged to predict the next word w_{t+1} through a softmax layer.

3.4 Training and Inference

In the training stage, we pre-construct the two kinds of visual graphs (i.e., semantic and spatial graphs) by exploiting the semantic and spatial relations among detected image regions as described in Sect. 3.2. Then, each graph is separately utilized to train one individual GCN-based encoder plus attention LSTM decoder. Note that the LSTM in decoder can be optimized with conventional cross entropy loss or the expected sentence-level reward loss as in [22,31].

At the inference time, we adopt a late fusion scheme to connect the two visual graphs in our designed GCN-LSTM architecture. Specifically, we linearly fuse the predicted word distributions from two decoders at each time step and pop

out the word with the maximum probability as the input word to both decoders at the next time step. The fused probability for each word w_i is calculated as:

$$\Pr(w_t = w_i) = \alpha \Pr_{sem}(w_t = w_i) + (1 - \alpha) \Pr_{spa}(w_t = w_i), \quad (8)$$

where α is the tradeoff parameter, $\Pr_{sem}(w_t = w_i)$ and $\Pr_{spa}(w_t = w_i)$ denotes the predicted probability for each word w_i from the decoder trained with semantic and spatial graph, respectively.

4 Experiments

We conducted the experiments and evaluated our proposed GCN-LSTM model on COCO captioning dataset (COCO) [21] for image captioning task. In addition, Visual Genome [16] is utilized to pre-train the object detector and semantic relation detector in our GCN-LSTM.

4.1 Datasets and Experimental Settings

COCO, is the most popular benchmark for image captioning, which contains 82,783 training images and 40,504 validation images. There are 5 human-annotated descriptions per image. As the annotations of the official testing set are not publicly available, we follow the widely used settings in [2,31] and take 113,287 images for training, 5K for validation and 5 K for testing. Similar to [13], we convert all the descriptions in training set to lower case and discard rare words which occur less than 5 times, resulting in the final vocabulary with 10,201 unique words in COCO dataset.

Visual Genome, is a large-scale image dataset for modeling the interactions/relationships between objects, which contains 108 K images with densely annotated objects, attributes, and relationships. To pre-train the object detector (i.e., Faster R-CNN in this work), we strictly follow the setting in [2], taking 98 K for training, 5K for validation and 5 K for testing. Note that as part of images (about 51K) in Visual Genome are also found in COCO, the split of Visual Genome is carefully selected to avoid contamination of the COCO validation and testing sets. Similar to [2], we perform extensive cleaning and filtering of training data, and train Faster R-CNN over the selected 1,600 object classes and 400 attributes classes. To pre-train the semantic relation detector, we adopt the same data split for training object detector. Moreover, we select the top-50 frequent predicates in training data and manually group them into 20 predicate/relation classes. The semantic relation detection model is thus trained over the 20 relation classes plus a non-relation class.

Features and Parameter Settings. Each word in the sentence is represented as "one-hot" vector (binary index vector in a vocabulary). For each image, we apply Faster R-CNN to detect objects within this image and select top $K = 36$ regions with highest detection confidences to represent the image. Each region is represented as the 2,048-dimensional output of pool5 layer after RoI pooling

from the Res4b22 feature map of Faster R-CNN in conjunction with ResNet-101 [11]. In the attention LSTM decoder, the size of word embedding D_s^1 is set as 1,000. The dimension of the hidden layer D_h in each LSTM is set as 1,000. The dimension of the hidden layer D_a for measuring attention distribution is set as 512. The tradeoff parameter α in Eq. (8) is empirically set as 0.7.

Implementation Details. We mainly implement our GCN-LSTM based on Caffe [12], which is one of widely adopted deep learning frameworks. The whole system is trained by Adam [14] optimizer. We set the initial learning rate as 0.0005 and the mini-batch size as 1,024. The maximum training iteration is set as 30 K iterations. For sentence generation in inference stage, we adopt the beam search strategy and set the beam size as 3.

Evaluation Metrics. We adopt five types of metrics: BLEU@N [28], METEOR [3], ROUGE-L [20], CIDEr-D [33] and SPICE [1]. All the metrics are computed by using the codes[1] released by COCO Evaluation Server [4].

Compared Approaches. We compared the following state-of-the-art methods: (1) **LSTM** [34] is the standard CNN plus RNN model which only injects image into LSTM at the initial time step. We directly extract results reported in [31]. (2) **SCST** [31] employs a modified visual attention mechanism of [37] for captioning. Moreover, a self-critical sequence training strategy is devised to train LSTM with expected sentence-level reward loss. (3) **ADP-ATT** [24] develops an adaptive attention based encoder-decoder model for automatically determining when to look (sentinel gate) and where to look (spatial attention). (4) **LSTM-A** [39] integrates semantic attributes into CNN plus RNN captioning model for boosting image captioning. (5) **Up-Down** [2] designs a combined bottom-up and top-down attention mechanism that enables region-level attention to be calculated. (6) **GCN-LSTM** is the proposal in this paper. Moreover, two slightly different settings of GCN-LSTM are named as GCN-LSTM$_{sem}$ and GCN-LSTM$_{spa}$ which are trained with only semantic graph and spatial graph, respectively.

Note that for fair comparison, all the baselines and our model adopt ResNet-101 as the basic architecture of image feature extractor. Moreover, results are reported for models optimized with both cross entropy loss or expected sentence-level reward loss. The sentence-level reward is measured with CIDEr-D score.

4.2 Performance Comparison and Experimental Analysis

Quantitative Analysis. Table 1 shows the performances of different models on COCO image captioning dataset. Overall, the results across six evaluation metrics optimized with cross-entropy loss and CIDEr-D score consistently indicate that our proposed GCN-LSTM achieves superior performances against other state-of-the-art techniques including non-attention models (LSTM, LSTM-A) and attention-based approach (SCST, ADP-ATT and Up-Down). In particular, the CIDEr-D and SPICE scores of our GCN-LSTM can achieve 117.1% and 21.1% optimized with cross-entropy loss, making the relative improvement over

[1] https://github.com/tylin/coco-caption.

Table 1. Performance of our GCN-LSTM and other state-of-the-art methods on COCO, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores. All values are reported as percentage (%).

	Cross-entropy loss						CIDEr-D score optimization					
	B@1	B@4	M	R	C	S	B@1	B@4	M	R	C	S
LSTM [34]	-	29.6	25.2	52.6	94.0	-	-	31.9	25.5	54.3	106.3	-
SCST [31]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
ADP-ATT [24]	74.2	33.2	26.6	-	108.5	-	-	-	-	-	-	-
LSTM-A [39]	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down [2]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM$_{spa}$	77.2	36.5	27.8	56.8	115.6	20.8	80.3	37.8	28.4	58.1	127.0	21.9
GCN-LSTM$_{sem}$	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
GCN-LSTM	**77.4**	**37.1**	**28.1**	**57.2**	**117.1**	**21.1**	**80.9**	**38.3**	**28.6**	**58.5**	**128.7**	**22.1**

the best competitor Up-Down by 3.2% and 3.9%, respectively, which is generally considered as a significant progress on this benchmark. As expected, the CIDEr-D and SPICE scores are boosted up to 128.7% and 22.1% when optimized with CIDEr-D score. LSTM-A exhibits better performance than LSTM, by further explicitly taking the high-level semantic information into account for encoding images. Moreover, SCST, ADP-ATT and Up-Down lead to a large performance boost over LSTM, which directly encodes image as one global representation. The results basically indicate the advantage of visual attention mechanism by learning to focus on the image regions that are most indicative to infer the next word. More specifically, Up-Down by enabling attention to be calculated at the level of objects, improves SCST and ADP-ATT. The performances of Up-Down are still lower than our GCN-LSTM$_{spa}$ and GCN-LSTM$_{sem}$ which additionally exploits spatial/semantic relations between objects for enriching region-level representations and eventually enhancing image captioning, respectively. In addition, by utilizing both spatial and semantic graphs in a late fusion manner, our GCN-LSTM further boosts up the performances.

Qualitative Analysis. Figure 5 shows a few image examples with the constructed semantic and spatial graphs, human-annotated ground truth sentences and sentences generated by three approaches, i.e., LSTM, Up-Down and our GCN-LSTM. From these exemplar results, it is easy to see that the three automatic methods can generate somewhat relevant and logically correct sentences, while our model GCN-LSTM can generate more descriptive sentence by enriching semantics with visual relationships in graphs to boost image captioning. For instance, compared to the same sentence segment "with a cake" in the sentences generated by LSTM and Up-Down for the first image, "eating a cake" in our GCN-LSTM depicts the image content more comprehensive, since the detected relation "eating" in semantic graph is encoded into relation-aware region-level features for guiding sentence generation.

Performance on COCO Online Testing Server. We also submitted our GCN-LSTM optimized with CIDEr-D score to online COCO testing server and

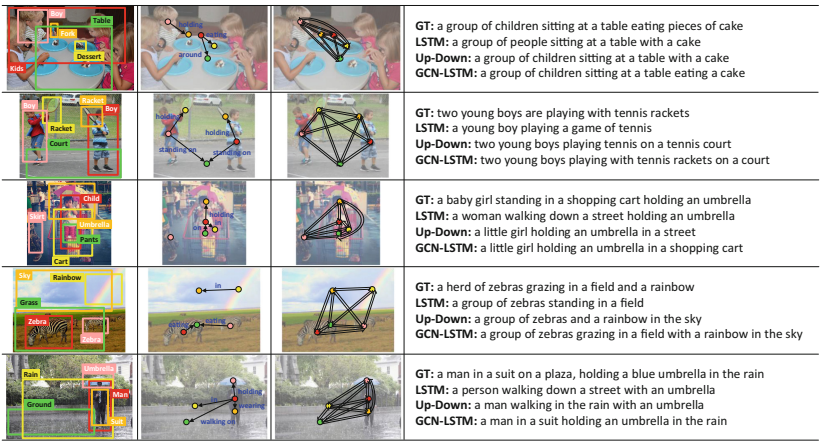

Fig. 5. Graphs and sentences generation results on COCO dataset. The semantic graph is constructed with semantic relations predicted by our semantic relation detection model. The spatial graph is constructed with spatial relations as defined in Fig. 4. The output sentences are generated by (1) Ground Truth (GT): One ground truth sentence, (2) LSTM, (3) Up-Down and (4) our GCN-LSTM.

Table 2. Leaderboard of the published state-of-the-art image captioning models on the online COCO testing server, where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr-D scores. All values are reported as percentage (%).

Model	B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
GCN-LSTM	**65.5**	**89.3**	**50.8**	**80.3**	**38.7**	**69.7**	**28.5**	**37.6**	**58.5**	**73.4**	**125.3**	**126.5**
Up-Down [2]	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
LSTM-A [39]	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
SCST [31]	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
G-RMI [22]	59.1	84.2	44.5	73.8	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
ADP-ATT [24]	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9

evaluated the performance on official testing set. Table 2 summarizes the performance Leaderboard on official testing image set with 5 (c5) and 40 (c40) reference captions. The latest top-5 performing methods which have been officially published are included in the table. Compared to the top performing methods on the leaderboard, our proposed GCN-LSTM achieves the best performances across all the evaluation metrics on both c5 and c40 testing sets.

Human Evaluation. To better understand how satisfactory are the sentences generated from different methods, we also conducted a human study to compare our GCN-LSTM against two approaches, i.e., LSTM and Up-Down. All of the three methods are optimized with CIDEr-D score. 12 evaluators are invited and a subset of 1 K images is randomly selected from testing set for the subjective evaluation. All the evaluators are organized into two groups. We show

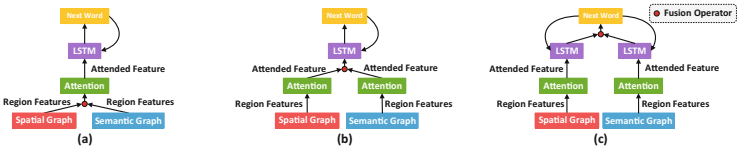

Fig. 6. Different schemes for fusing spatial and semantic graphs in GCN-LSTM: (a) Early fusion before attention module, (b) Early fusion after attention module and (c) Late fusion. The fusion operator could be concatenation or summation.

the first group all the three sentences generated by each approach plus five human-annotated sentences and ask them the question: Do the systems produce captions resembling human-generated sentences? In contrast, we show the second group once only one sentence generated by different approach or human annotation (Human) and they are asked: Can you determine whether the given sentence has been generated by a system or by a human being? From evaluators' responses, we calculate two metrics: (1) M1: percentage of captions that are evaluated as better or equal to human caption; (2) M2: percentage of captions that pass the Turing Test. The results of M1 are 74.2%, 70.3%, 50.1% for GCN-LSTM, Up-Down and LSTM. For the M2 metric, the results of Human, GCN-LSTM, Up-Down and LSTM are 92.6%, 82.1%, 78.5% and 57.8%. Overall, our GCN-LSTM is clearly the winner in terms of two criteria.

Effect of Fusion Scheme. There are generally two directions for fusing semantic and spatial graphs in GCN-LSTM. One is to perform early fusion scheme by concatenating each pair of region features from graphs before attention module or the attended features from graphs after attention module. The other is our adopted late fusion scheme to linearly fuse the predicted word distributions from two decoders. Figure 6 depicts the three fusion schemes. We compare the performances of our GCN-LSTM in the three fusion schemes (with cross-entropy loss). The results are 116.4%, 116.6% and 117.1% in CIDEr-D metric for early fusion before/after attention module and late fusion, respectively, which indicate that the adopted late fusion scheme outperforms other two early fusion schemes.

Effect of the Tradeoff Parameter α. To clarify the effect of the tradeoff parameter α in Eq. (8), we illustrate the performance curves over three evaluation metrics with a different tradeoff parameter in Fig. 7. As shown in the figure, we

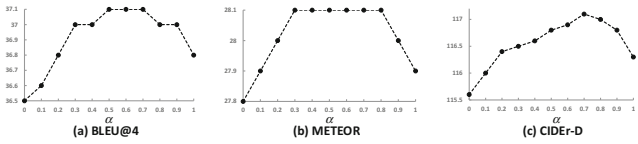

Fig. 7. The effect of the tradeoff parameter α in our GCN-LSTM with cross-entropy loss over (a) BLEU@4 (%), (b) METEOR (%) and (c) CIDEr-D (%) on COCO.

can see that all performance curves are generally like the "∧" shapes when α varies in a range from 0 to 1. The best performance is achieved when α is about 0.7. This proves that it is reasonable to exploit both semantic and spatial relations between objects for boosting image captioning.

5 Conclusions

We have presented Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) architecture, which explores visual relationship for boosting image captioning. Particularly, we study the problem from the viewpoint of modeling mutual interactions between objects/regions to enrich region-level representations that are feed into sentence decoder. To verify our claim, we have built two kinds of visual relationships, i.e., semantic and spatial correlations, on the detected regions, and devised Graph Convolutions on the region-level representations with visual relationships to learn more powerful representations. Such relation-aware region-level representations are then input into attention LSTM for sentence generation. Extensive experiments conducted on COCO image captioning dataset validate our proposal and analysis. More remarkably, we achieve new state-of-the-art performances on this dataset. One possible future direction would be to generalize relationship modeling and utilization to other vision tasks.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
3. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop (2005)
4. Chen, X., et al.: Microsoft COCO captions: data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
5. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: CVPR (2017)
6. Divvala, S.K., Farhadi, A., Guestrin, C.: Learning everything about anything: webly-supervised visual concept learning. In: CVPR (2014)
7. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR (2015)
8. Fu, K., Jin, J., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans. PAMI **39**, 2321–2334 (2017)
9. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR (2008)
10. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. IJCV **80**, 300–316 (2008)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. Jia, Y., et al.: Caffe: Convolutional architecture for fast feature embedding. In: MM (2014)
13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
14. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
16. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. IJCV (2017)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
18. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Jointly localizing and describing events for dense video captioning. In: CVPR (2018)
19. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)
20. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: ACL Workshop (2004)
21. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
22. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Optimization of image description metrics using policy gradient methods. In: ICCV (2017)
23. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
24. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: CVPR (2017)
25. Marcheggiani, D., Titov, I.: Encoding sentences with graph convolutional networks for semantic role labeling. In: EMNLP (2017)
26. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: CVPR (2016)
27. Pan, Y., Yao, T., Li, H., Mei, T.: Video captioning with transferred semantic attributes. In: CVPR (2017)
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL (2002)
29. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: ICCV (2017)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015)
31. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)
32. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
33. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR (2015)
34. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR (2015)

35. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: CVPR (2016)
36. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR (2017)
37. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML (2015)
38. Yao, T., Pan, Y., Li, Y., Mei, T.: Incorporating copying mechanism in image captioning for learning novel objects. In: CVPR (2017)
39. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: ICCV (2017)
40. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016)