Pattern Recognition 98 (2020) 107075

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning visual relationship and context-aware attention for image captioning

Junbo Wang[a,c,*], Wei Wang[a,c,*], Liang Wang[a,b,c], Zhiyong Wang[d], David Dagan Feng[d], Tieniu Tan[a,b,c]

[a] Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) China
[b] Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), CASIA, China
[c] University of Chinese Academy of Sciences (UCAS), China
[d] School of Information Technologies, The University of Sydney, Australia

ARTICLE INFO

Article history:
Received 26 September 2018
Revised 27 September 2019
Accepted 7 October 2019
Available online 8 October 2019

Keywords:
Image captioning
Relational reasoning
Context-aware attention

ABSTRACT

Image captioning which automatically generates natural language descriptions for images has attracted lots of research attentions and there have been substantial progresses with attention based captioning methods. However, most attention-based image captioning methods focus on extracting visual information in regions of interest for sentence generation and usually ignore the relational reasoning among those regions of interest in an image. Moreover, these methods do not take into account previously attended regions which can be used to guide the subsequent attention selection. In this paper, we propose a novel method to implicitly model the relationship among regions of interest in an image with a graph neural network, as well as a novel context-aware attention mechanism to guide attention selection by fully memorizing previously attended visual content. Compared with the existing attention-based image captioning methods, ours can not only learn relation-aware visual representations for image captioning, but also consider historical context information on previous attention. We perform extensive experiments on two public benchmark datasets: MS COCO and Flickr30K, and the experimental results indicate that our proposed method is able to outperform various state-of-the-art methods in terms of the widely used evaluation metrics.

1. Introduction

Image captioning [1–3], which automatically generates natural language descriptions for images, has a wide range of applications, such as image retrieval, aiding the visually impaired, and intelligent human computer interaction. For decades, it has been a challenging cross-disciplinary task involving both computer vision and natural language processing.

Recently, deep learning techniques such as Convolutional Neural Network (CNN) [4,5] and Recurrent Neural Network (RNN) [6,7] have significantly contributed to the great progresses in image captioning [8–10]. In particular, various visual attention-based encoder-decoder models have been widely explored for image captioning [11–13] with great success by emphasizing visually important content. However, these methods often have the following two limitations. First, while specific regions or objects of interests in an image are attended during sentence generation, the relationship among those regions or objects has not yet been explored. For example, to caption an image with human-annotated description "the man is walking a herd of sheep on the road through a town", a captioning method needs to figure out the relationship among visual objects in the image, i.e., the relationship "walking" between "man" and "a herd of sheep" and the relationship "on" between "a herd of sheep" and "road". Second, most current attention-based image captioning methods focus on objects/regions most relevant to the word being generated at each time step, and ignore what has been attended to at previous time steps. As a result, these models may attend to the same region in an image at multiple time steps, which could compromise the effectiveness of the captioning method.

Based on the above observations, we leverage a graph neural network (GNN [14]) to implicitly model the visual relationship between objects or regions in an image and propose a visual context-

* Corresponding authors at: Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) China.
E-mail addresses: junbo.wang@nlpr.ia.ac.cn (J. Wang), wangwei@nlpr.ia.ac.cn (W. Wang), wangliang@nlpr.ia.ac.cn (L. Wang), zhiyong.wang@sydney.edu.au (Z. Wang), dagan.feng@sydney.edu.au (D.D. Feng), tnt@nlpr.ia.ac.cn (T. Tan).

https://doi.org/10.1016/j.patcog.2019.107075
0031-3203/

aware attention mechanism to guide sentence generation with previously attended content. In particular, we first utilize a deep CNN to extract visual representation of an input image, and consider each region of interest as a node and build a relationship graph where all the nodes are fully-connected in an undirected way. The GNN [14] propagates messages along all edges in a recurrent manner and outputs all representations corresponding to the nodes in the graph, which can be viewed as implicit relation-aware visual representations among objects in the image. Then our context-aware attention model will attend to the learned relationship representations at each time step. To memorize what has been attended to, we use a Long Short Term Memory (LSTM) to keep track of previously attended visual content and fuse the attention weight produced by our visual attention model at the current time step with the attention weight produced at the previous time steps for a joint attention model. Finally, we employ a LSTM-based language model to predict next word given previously generated word and relation-aware visual representations selected by our context-aware attention model.

In summary, main contributions of our work are as follows:

- We propose to implicitly model the relationship among the objects/regions in an image with a GNN, which takes into account the visual relationship among regions of interest for better representation of the visual content in the image.
- We propose a novel visual context-aware attention model to select salient visual information at each time step, which utilizes a contextual LSTM to keep track of previously attended visual information and combine the attention weight produced by our attention model at the current time step with the attention weight produced at the previous time step.
- We conduct extensive experiments to quantitatively evaluate our proposed method on two public benchmark datasets: MS COCO and Flickr30K. Experimental results demonstrate that our proposed method performs much better than other state-of-the-art methods.

The remainder of this paper is organized as follows. In Section 2, we first review some most relevant studies: image captioning and graph neural network. In Section 3, we introduce the overall framework of our proposed method and detail each component in the framework. In Section 4, we describe the experimental datasets, training setups, evaluation metrics, quantitative and qualitative analysis. In Section 5, we draw our conclusions and future work on this topic.

2. Related work

In this section, we review two types of studies most relevant to our work: image captioning and graph neural network.

2.1. Image captioning

The recent work on image captioning can be grouped into three categories: template-based methods, retrieval-based methods and neural network-based methods. The template-based methods [15–18] first detect key visual concepts (e.g., objects and attributes) from images by utilizing object detection and attribute classification methods. According to predefined language templates, these methods split a sentence into several parts (e.g., subject, verb and object). Finally, these methods align the detected visual concepts with the parts in a language template via statistical methods (e.g., CRF [16] and HMM [17]). Since these methods highly rely on predefined language templates, they can only generate syntactically correct sentences at the loss of the flexibility of natural language.

The retrieval-based methods [19–22] usually measure the similarity between an input image and external sentences or the similarity between the input image and other visually similar images. Based on the similarity, these methods can choose most semantically similar sentences from an external sentence pool or candidate sentences extracted from those visually similar images. As a result, these methods can generate human-level sentences as all the sentences were manually produced by humans. However, these methods are difficult to be transferred to different datasets and cannot generate novel sentences for images.

The neural network-based methods [8,23,24] utilize deep neural networks to exploit conditional probability distribution given the visual content and generated words. Inspired by the success of encoder-decoder models in neural machine translation [6,7], these methods consider image captioning as a translation task (bridging source image to target language). For example, Vinyals et al. [8] employ a deep CNN to encode an input image into a static vector and utilize a LSTM-based language model to decode a sentence based on the encoded vector. Similarly, Mao et al. [25] first use a deep convolutional network to encode an input image and employ a RNN-based language model to encode previously generated words, then propose a multimodal model to combine both visual and textual information to predict next word. Karpathy et al. [24] also propose a multimodal recurrent neural network model to align information of two modalities as well as simultaneously locate the key objects in the generated sentence. Donahue et al. [23] and Jia et al. [26] both explore different ways of combining visual information with LSTM block to guide sentence generation. However, the encoded static vector in the abovementioned methods is not sufficient to represent the whole image due to the missing objects. Inspired by the success of attention mechanism in natural language processing [27] and computer vision [28,29], Xu et al. [11] propose a visual attention model to select the most relevant region representation for generating each word during sentence generation, instead of using a global static vector. As a result, this method can generate a sentence according to different visual content at each time step. However, the visual attention model has to attend to visual content even when generating non-visual words (e.g., “a”, “the” and “of”). Therefore, Lu et al. [13] propose a novel adaptive attention model to determine whether to attend to the image or to the visual sentinel to extract meaningful information for sentence generation. Different from these visual attention models, Zhou et al. [30] propose a text-conditional attention model to allow the caption generator to attend to certain visual content given previously generated words. You et al. [31] and Wu et al. [32] propose a semantic attention model to selectively attend to semantic concept proposals and incorporate them into the input and output of the LSTM-based language model via the top-down and bottom-up computation. However, most existing works try to optimize the likelihood of the next ground-truth word using back-propagation, which leads to the exposure bias between training and testing. To address this problem, recent works [33,34] employ policy-gradient methods to directly optimize non-differentiable metrics for the task. Some researchers [35,36] also replace RNN-based language models with CNN-based language models to address the inefficiency of LSTM across time during sentence generation. Furthermore, in order to obtain better image representations for image captioning, previous works [9,10] first generate several object proposals and extract corresponding features of these object proposals for further processing. Recent works [37,38] also leverage visual relationships to generate region/image captioning, which detect visual relationship classes based on visual objects explicitly detected by the Faster R-CNN object proposal network pre-trained on the Visual Genome [39] dataset, while our work learns implicit visual relationships on image regions of interest on the COCO/Flickr dataset, which does not need pre-defined relationship classes and explicit object detections.

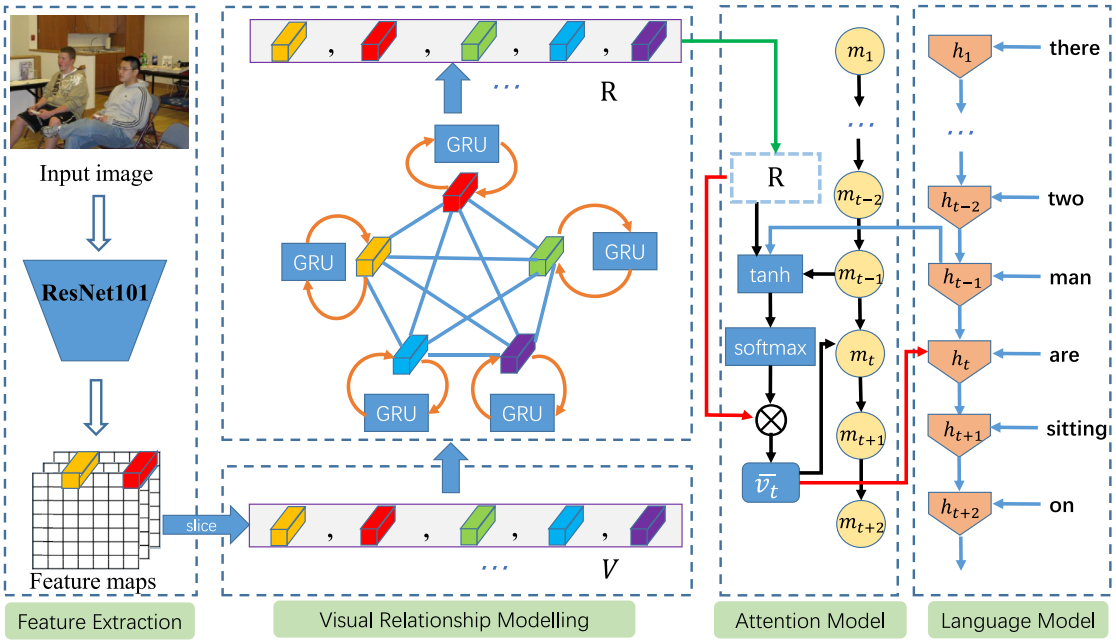

Fig. 1. The overall framework of our proposed image captioning method which consists of four components: CNN-based deep feature extraction, graph-based visual relationship modelling, visual context-aware attention model and LSTM-based language model.

2.2. Graph Neural Network

To apply neural networks to graph data, a GNN employs feed-forward neural networks to all nodes of the graph in a recurrent way. At each time step, the GNN takes previous hidden state of each node and the messages from its adjacent nodes as input to dynamically update the current hidden state of each node. In [14], the GNN employs multi-layer perceptrons (MLP) to update the hidden state of each node. However, to ensure the gradient-descent strategy based learning algorithm converge, their contraction map assumption has trouble to propagate information across a long range in a graph. To address the problem, Li et al. [40] propose Gated Graph Neural Network (GGNN) for some graph data based learning tasks where gated recurrent units are employed to update the hidden state of a node in the graph by using the backpropagation through time strategy to compute gradients. Other works [41,42] apply convolutional neural networks to the graph domain by encoding both local graph structure and features of nodes for the classification of graph data. Wang et al. [43] perform similarity relationship and spatial-temporal relationship reasoning with graph convolutional networks for human action recognition. Petarn et al. [44] leverage masked self-attentional layers in the graph attention network to address the shortcoming of these graph convolution based methods. Our work utilizes a GNN model to explore the implicit visual relationship among the objects/regions of interest in an image.

3. Our proposed method

The overall framework of our image captioning system is illustrated in Fig. 1. It consists of a deep CNN to extract image features, a GNN model to learn the implicit visual relationship among the visual objects or regions in an image, a visual context-aware attention model to select important relationship representations, and a LSTM-based language model to generate sentences.

Given an image I, we employ the widely-used CNN architecture ResNet101 [45] pretrained on the ImageNet classification task to extract nonlinear activations from the last convolutional layer as image representations, which can be denoted as $V = \{v_1, v_2, \cdots, v_n | v_i \in \mathbb{R}^m\}$. Based on the image representations corresponding to different spatial locations, we utilize a GNN model f_{gnn} to explore implicit relationship between the visual objects in the image. The GNN model takes each spatial representation to initialize each node in the graph and recurrently updates each node information by utilizing the hidden presentations from other nodes to obtain the implicit relation-aware visual representations $R = \{r_1, r_2, \cdots, r_n | r_i \in \mathbb{R}^m\}$. The generated implicit relation-aware visual representations R are forwarded into a context-aware attention model f_{att}. Different from the existing visual attention models, our context-aware attention model employs a LSTM to record previously attended visual information at each time step, which helps guide the future selection on the unexplored visual information in an inhibition-of-return way. After that, a LSTM-based language model f_{lstm} takes previous hidden state h_{t-1}, previously generated word embedding x_t and the outputs $\bar{v}_t$ of the context-aware attention model as input, and outputs the current hidden state h_t to predict the next word. The main working flow of our image captioning method is shown in the following equations:

$$V = CNN(I), \tag{1}$$

$$R = f_{gnn}(V), \tag{2}$$

$$\bar{v}_t = f_{att}(R, h_{t-1}, p_{t-1}), \tag{3}$$

$$h_t = f_{lstm}(h_{t-1}, x_t, \bar{v}_t), \tag{4}$$

$$s_t = \arg\max_s softmax(W_o h_t + b_o), \tag{5}$$

where t denotes the time step, s_t denotes the predicted word according to the maximum softmax probability, W_o and b_o are the weight and bias to be learned respectively. The hidden state h_0 is initialized with zero. Eqs. (3)–(5) are recursively applied, f_{gnn}, f_{att} and f_{lstm} will be discussed in the following sections.

3.1. Graph-based visual relationship modelling

The GNN models data structure and representation in a graph, which has made remarkable success on various graph data based

learning tasks. Inspired by GGNN [40] which learns the representation of a graph to predict node or graph-level output, we extend it to explore the implicit relationship among the visual objects in images. In the GNN model, we instantiate a graph G for each image that consists of N nodes corresponding to spatial locations of the image representation derived from a deep CNN. In order to fully capture the relationships between these nodes, we employ a fully-connected graph and learn edge strengths/weights between nodes. The edge strengths/weights between nodes form an adjacency matrix A, which denote the probabilities of the relationships existing between any two graph nodes. Without the learning of edge weights, each edge weight $A_{i,j}$ except for the diagonal element in the adjacency matrix is one. With the learning of edge weights, each edge weight $A_{i,j}$ between two nodes v_i, v_j in the graph can be defined as:

$$A_{i,j} = \sigma\left(f_{edge}\left(\left|v_i - v_j\right|\right)\right) \tag{6}$$

where f_{edge} is a convolutional layer with kernel size 1, followed by a sigmoid function. The f_{edge} takes the absolute difference between node features as input, which satisfies the symmetry property [46].

To reduce the dimension of image representations and initialize the hidden state of each node in the graph, we apply non-linear transformation to the image representations V and use the transformed vector to initialize the hidden state of each node:

$$\widetilde{v}_a^t = \varphi(W_a v_a + b_a), \tag{7}$$

$$h_a^0 = \beta\left(\widetilde{v}_a^t\right), \tag{8}$$

where W_a and b_a are the weight and bias to be learned respectively, $v_a \in V$ is the feature vector corresponding to each spatial location in the image, h_a^0 denotes the initial hidden state of each node a in the graph, φ and β are the non-linear activation functions (e.g., hyperbolic tangent function Tanh and rectified linear unit Relu).

At each time step t, the incoming messages of each node a are collected from the hidden states of its adjacent nodes $\{d|\forall a \in G, (d, a) \in \mathrm{B}\}$:

$$x_a^t = \sum_{(d,a)\in \mathrm{B}} W_g h_d^{t-1} + b_g, \tag{9}$$

where W_g and b_g are the shared weight and bias to be learned across all nodes respectively, and B denotes the collection of adjacent nodes, which can be obtained from the adjacency matrix A.

After aggregating the incoming messages for each node, the GNN employs Gated Recurrent Unit (GRU) which contains a reset gate r and a update gate z to update the hidden state of each node as follows:

$$z_a^t = \sigma\left(W_z x_a^t + U_z h_a^{t-1} + b_z\right), \tag{10}$$

$$r_a^t = \sigma\left(W_r x_a^t + U_r h_a^{t-1} + b_r\right), \tag{11}$$

$$\widetilde{h}_a^t = \phi\left(W_h x_a^t + U_h\left(r_a^t \odot h_a^{t-1}\right) + b_h\right), \tag{12}$$

$$h_a^t = \left(1 - z_a^t\right) \odot h_a^{t-1} + z_a^t \odot \widetilde{h}_a^t, \tag{13}$$

where the default operation between matrices is matrix multiplication, $\odot$ denotes an element-wise multiplication, W and U denote the shared weights to be learned, b denotes the bias term, σ denotes the element-wise logistic sigmoid function, and ϕ denotes hyperbolic tangent function tanh. The reset gate r and the update gate z selectively control the influence of information from previous hidden state and current hidden state. Note that the hidden states of all nodes in the graph are updated synchronously. We recurrently update the hidden state of each node for T time steps and extract node-level outputs of the GNN as the implicit relation-aware visual representations R for the following visual context-aware attention model.

3.2. Visual context-aware attention model

Given the previous hidden state h_{t-1} of the LSTM-based language model, the previously attended visual information p_{t-1} and the implicit relationship representations R from the GNN, the initial normalized attention weights a_t for the visual signal R can be obtained through a single layer neural network followed by a softmax function:

$$z_t = w_{att}^{\mathrm{T}} \tanh\left(U_{att} R + W_{att} h_{t-1} + M_{att} p_{t-1} + b_{att}\right), \tag{14}$$

$$a_t = softmax(z_t), \tag{15}$$

where U_{att}, W_{att} and M_{att} are the shared weights to be learned, b_{att} denotes the bias term. In this step, the implicit relation-aware visual representations R are forwarded into the attention model. As a result, the attention model can attend to the implicit visual relationship at each time step. Furthermore, we design an interpolation gate k_t to fuse the current normalized weight a_t with previously produced weight $\bar{a}_{t-1}$.

$$k_t = \sigma(W_k h_{t-1} + b_k), \tag{16}$$

$$\bar{a}_t = k_t a_t + (1 - k_t)\bar{a}_{t-1}, \tag{17}$$

where W_k and b_k are the shared weight and bias term to be learned respectively, σ denotes the element-wise logistic sigmoid function. If the gate k_t is zero, the current normalized weight is entirely ignored, and the previously produced weight is used. On the contrary, if the gate k_t is one, the previously produced weight is ignored, and the current normalized weight is applied to select suitable visual information. The attended visual signal $\bar{v}_t$ is denoted as a linear combination of all relation-aware visual representations:

$$\bar{v}_t = \sum_{i=1}^{n} \bar{a}_{i,t}^r (R)_i, \tag{18}$$

After obtaining the attended visual information, we forward it into a LSTM model which memorizes the visual information selected by our attention model. As a result, the context information from the LSTM model can be utilized to guide the attention weight selection at next time step:

$$p_t = q_{lstm}(p_{t-1}, \bar{v}_t). \tag{19}$$

3.3. LSTM-based language model

To model sentence generation, we employ a variant of LSTM [47] which has achieved great success in image captioning. Different from previous image captioning models, we design an adaptive gate g_t to control whether visual signal can be fed into the iteration of LSTM. The proposed visual gated LSTM extends the basic LSTM which contains a memory cell m_t and three input gates (i.e., input gate i_t, forget gate f_t and output gate o_t) with additional visual gate unit. The inputs to the visual gated LSTM include the word embedding x_t, the previous hidden state h_{t-1} and the attended visual signal $\bar{v}_t$. The iteration of LSTM at each time step t can be formulated as follows:

$$g_t = \sigma(W_g x_t + U_g h_{t-1} + b_g), \tag{20}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + g_t \odot M_i \bar{v}_t + b_i), \tag{21}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + g_t \odot M_f \bar{v}_t + b_f\right), \tag{22}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + g_t \odot M_o \bar{v}_t + b_o), \tag{23}$$

$$\widetilde{m}_t = \phi(W_c x_t + U_c h_{t-1} + g_t \odot M_c \bar{v}_t + b_c), \tag{24}$$

$$m_t = i_t \odot \widetilde{m}_t + f_t \odot m_{t-1}, \tag{25}$$

$$h_t = o_t \odot \phi(m_t), \tag{26}$$

where the default operation between matrices is matrix multiplication, $\odot$ denotes an element-wise multiplication, W, U, and M denote the shared weight matrices to be learned, and b denotes the bias term. $\tilde{m}_t$ is the input to the memory cell m_t, which is gated by the input gate i_t. σ denotes the element-wise logistic sigmoid function, and ϕ denotes hyperbolic tangent function tanh.

3.4. Model learning

Based on the hidden state of LSTM-based language model at each timestep, we employ a non-linear softmax layer to predict the next word's probability distribution over the whole vocabulary:

$$\rho_t = softmax\left(U_\rho h_t + b_\rho\right), \tag{27}$$

where U_ρ and b_ρ denote the parameters to be learned.

During training, the optimal sentence corresponding to the input image can be generated by maximizing the probability of sentences via chain rule. Assuming that there are N image-description training pairs (x^i, y^i) in the training dataset, where each sentence y^i has a variable length t_i. We define the overall loss function as the averaged log-likelihood over the whole training dataset plus a regularization term:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{t_i}\log \rho\left(y_j^i | y_{1:j-1}^i, x^i, \theta\right) + \lambda\|\theta\|_2^2, \tag{28}$$

where y_j^i is a one-hot encoding vector corresponding to the input word, θ is model parameters to be learned, and λ denotes the regularization coefficient. We can use stochastic gradient descent to optimize the above loss function. During testing, we can recursively sample y_t based on the probability distribution ρ_t until meeting the end symbol of the vocabulary.

4. Experimental results and discussions

4.1. Datasets

We compare our proposed model with the state-of-the-art methods on two public benchmark datasets, Microsoft COCO [48] and Flickr30k [49].

Microsoft COCO consists of 82,783, 40,504 and 40,775 images for training, validation and testing respectively. It is currently the largest image captioning dataset. Each image is labelled with at least five captions in the dataset. Compared with Ficker30k, this dataset is more challenging since the images contain multiple objects in their natural context. As there are no available ground truth captions for the test set, we follow the widely used data split [24] for this dataset: 5000 images for validation, 5000 images for testing and other images for training.

Flickr30k consists of 158,915 crowd-sourced captions and 31,783 images collected from Flickr. This dataset extends the previous Flickr8k dataset and mainly describes everyday activities and events on humans. Each image has five reference captions in the dataset. To make fair comparison with existing studies, we employ the publicly available split [24]: 29,783 images are used for training, 1000 images for validation, and 1000 images for testing.

4.2. Experimental settings

Data preprocessing. In the experiments, we apply the standard preprocessing practice to the images and captions.

For captions, we convert each sentence to lower case and discard all the non-alphabetic characters. We drop those words that occur less than five times **in MS COCO** or three times **in Flick30k**, resulting in a vocabulary with size 10,478 and 7652 in MS COCO and Flickr30k respectively. If a word is not in the vocabulary, we set it as an unknown token $<\text{UNK}>$. For modelling convenience, we add a start token $<\text{Start}>$ and an end token $<\text{End}>$ to the vocabulary. During testing, we set the maximum allowed sentence length as 30.

For images, we encode them using the spatial feature outputs of the last convolutional layer of ResNet-101 [45]. After forwarding the images to the ResNet-101, we employ spatially adaptive average pooling used in [33] to make the output size of all images same. Therefore, the final output size of the last convolutional layer of ResNet-101 is $10 \times 10 \times 2048$, resulting in the 100 spatial location indexes over the input image.

Training and testing details. In the experiments, we set the hidden size of LSTM, the image feature size and word embedding size all to 512. To decrease the number of model parameters, we set the hidden size of each attention layer to 256. We employ Adam optimizer with the initial learning rate 5e-4. We set batch size as 64 and maximum epoch number as 80. To avoid overfitting, we employ dropout with rate 0.5 and early stopping if the validation CIDEr [50] score does not increase over the last 10 epochs. During training, we first fix the deep CNN part for training up to 30 epoches and then finetune the deep CNN part by annealing the learning rate by a factor of 0.8 every three epochs. During testing, we forward the start token or previously generated token to the trained model to sample next word until the end token is reached. Similar to existing image captioning models, we use beam search strategy with size 5. Even though we find that beam search with length normalization can improve performance, we do not use length normalization in all experiments to keep comparisons fair.

4.3. Evaluation metrics

To quantitatively evaluate the performance of our proposed method, four commonly used metrics, namely BLEU [51], Meteor [52], Rouge-L [53] and CIDEr [50]), are used to evaluate the quality of generated sentences. All these metrics measure the consistency of n-grams between generated sentences and reference sentences. To make fair comparisons with the existing image captioning methods, we utilize the publicly available implementation evaluation code released by MS COCO Evaluation Server [48] to test the performance.

4.4. Compared methods

To demonstrate the effectiveness of our proposed method, we compared the following state-of-the-art methods:

(1) NIC [8]: NIC injects image features derived from the fully-connected layer of a deep CNN into the first time step of the LSTM-based language model. We directly cite the results reported in [54].
(2) LRCN [23]: LRCN considers two stacked LSTM as a language model which takes previously generated word and global image feature derived from the fully-connected layer of a deep CNN as input at each timestep.
(3) DeepVS [24]: DeepVS first learns a structured objective that aligns two modalities (image regions and sentences) through a multimodal embedding, then utilizes a multimodal recurrent neural network to generate sentences corresponding to image regions based on the learned modal alignments.
(4) Soft-Att [11] and Hard-Att [11]: Soft-Att and Hard-Att select some regional representations derived from the last convolutional layer of a deep CNN and use the LSTM-based language model to decode each word at each timestep conditioned on the selected representations.
(5) ATT-FCN [31]: ATT-FCN first detects key attributes in an image, then takes the global image feature and the detected

attributes as input and fuses them into the hidden state of the LSTM-based language model at each timestep.

(6) G-LSTM [26]: G-LSTM incorporates extra semantic information obtained from retrieval-based guidance, semantic embedding guidance and image-based guidance into the LSTM-based language model to generate captions.

(7) ERD [12]: ERD conducts a fixed number of review steps including attentive input reviewer and attentive output reviewer on the encoder to generate multiple thought vectors.

(8) Sentence-Condition (SC) [30]: SC leverages previously generated text to guide the model to focus on certain image features and injects the attended features into the LSTM-based language model at each timestep.

(9) MSM [54]: MSM integrates the inter-attribute correlations into multiple instance learning method and explores different ways of injecting the detected attributes and image representations into the LSTM-based language model.

(10) Adap [13]: Adap is a novel adaptive attention model which determines whether to attend to the image feature or not as the prediction of some words does not need visual signal.

(11) *Att2in** [33]: *Att2in** employs an improved attention model for sentence generation and leverages a self-critical sequence training algorithm to optimize non-differentiable NLP metrics to boost the model performances. To keep a fair comparison, we only cite the results under the same optimization objective.

(12) SCA [55]: SCA extends previous Soft-Att [11] with channel-wise attention in multi-layer feature maps, which can dynamically modulate visual context across spatial, channel-wise and multi-layer dimensions.

(13) UD-Base* [56]: UD-Base* first detects key image regions via a Faster R-CNN model, then use a top-down attention model to select the regions. To keep a fair comparison, we only cite the results of the proposed model under the same image features and training objective.

(14) Convcap [35]: Convcap employs a CNN-based decoder as a language model to generate sentences. The CNN-based decoder is mainly implemented by multi-layer masked convolutions.

(15) AED-AR [57]: AED-AR attempts to regularize the transition dynamics of the LSTM-based language model with an auto-reconstructor network.

(16) WICG [58]: WICG explores different ways of incorporating image features into the language model and demonstrates that merging image features in a subsequent stage is effective.

(17) HCVSA [59]: HCVSA utilizes a bidirectional Grid LSTM to learn complex spatial patterns in the image context and employs a two-layer bidirectional LSTM to generate the global sentence.

(18) Our_A_R_L is proposed in this paper, and other variant models are also explored. Our Baseline employs the attention model as [11] and the LSTM-based language model described in Section 3.3, Our_A incorporates context-aware attention model to the baseline, Our_R incorporates graph-based visual relationship without the learning of edge weights to the baseline, Our_R_L incorporates graph-based visual relationship with the learning of edge weights to the baseline, and Our_A_R employs both attention context-aware attention model and graph-based relationship information.

4.5. Quantitative analysis

Table 1 shows the performance of compared methods and ours on the test split of MS COCO. Overall, the results across seven evaluation metrics consistently indicate that our proposed Our_A_R and Our_A_R_L achieve better performances than other eighteen state-of-the-art methods. In particular, Our_A_R_L can achieve 35.8 and 111.3 in the BLEU@4 and CIDEr respectively, making the relative improvement over the recently state-of-the-art attention-based methods (Att2in [33] and Adap [13]) by 14.4% / 9.9% and 7.8% / 2.6% respectively. Note that our Baseline which employs the same attention mechanism as [11] and the LSTM-based language model described in Section 3.3 also achieves better or comparable performance than some state-of-the-art methods (e.g., SC [30] and MSM [54]) due to the powerful ability of variant LSTM. By additionally incorporating context-aware attention mechanism and graph-based visual relationship to Baseline respectively, both Our_A and Our_R can achieve further performance improvement in terms of all evaluation metrics compared with the implemented Baseline. The improvement of Our_A and Our_R over Baseline by 3.1% and 4.2% respectively in the CIDEr metric indicates that our proposed attention model and graph-based relationship model are helpful for image captioning. When utilizing both attention context-aware visual attention mechanism and graph-based relationship information, our proposed Our_A_R can significantly improve captioning performance from 32.3/101.6 to 35.2/109.4 in terms of BLEU@4 and CIDEr, respectively. After using learning-based edges for visual relationship modelling, our model can further achieve better results in most evaluation metrics. These results indicate that exploiting context-aware visual attention mechanism and building graph-based relationship model are complementary for improving image captioning performance.

The performance comparison in terms of seven evaluation metrics on the test split of the Flickr30k dataset is summarized in Table 2. The evaluation scores on Flickr30k are much lower than those on MS COCO, due to the small number of training samples including visual and textual clues in the dataset. Similarly, our proposed Our_A_R and Our_A_R_L consistently outperform other state-of-the-art methods in terms of all evaluation metrics. In particular, our proposed Our_A_R_L achieves 27.7 and 57.4 in the BLEU@4 and CIDEr, respectively, making the relative improvement over the best competitor Adap [13] by 10.4% and 8.1% respectively. Similar to the observations on MS COCO, our proposed Our_A and Our_R perform much better than Baseline by further taking context-aware visual attention mechanism and graph-based relationship model into account for image captioning respectively. In addition, further improvement is achieved with Our_A_R where both context-aware visual attention model and graph-based relationship model are utilized. When using learning-based edges for our models (Our_R_L and Our_A_R_L), the performances can be further boosted in terms of most evaluation metrics.

4.6. Qualitative analysis

To better visualize and understand the visual relationships, we plot the relationship probability matrices (also called relationship adjacency matrix) of two test images in the Fig. 2(a). The learned edge strengths/weights denote the probabilities of the relationships existing between any two graph nodes, and the sparse probability matrix means the sparse relationships between the objects in the images. To better illustrate the learned relationships, we then plot the attention weight distribution over 100 graph nodes when generating three subject-relation-object words in Fig. 2(b), e.g., zebra-standing-snow and man-riding-wave, and show the strong relationships corresponding to the mostly attended three graph nodes in Fig. 2(c). Overall, it can be seen that our model can capture rich visual relationships consistent with human perception on the test images. From the three subject-relation-object words in Fig. 2(b,c), e.g., zebra-standing-snow and man-riding-wave, we can see that the nodes and their relationships corresponding to

Table 1
The performance comparison with eighteen state-of-the-art methods on the MS COCO dataset. The results of ablated models (Baseline, Our_A, Our_R, Our_R_L and Our_A_R) and our full model (Our_A_R_L) are shown at the bottom of the table.

Method	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
NIC [8]	66.6	45.1	30.4	20.3	-	-	-
LRCN [23]	69.7	51.9	38.0	27.8	22.9	50.8	83.7
DeepVS [24]	62.5	45.0	32.1	23.0	19.5	-	66.0
Soft-Att [11]	71.8	50.4	35.7	25.0	23.0	-	-
Hard-Att [11]	70.7	49.2	34.4	24.3	23.9	-	-
ATT-FCN [31]	70.9	53.7	40.2	30.4	24.3	-	-
ERD [12]	-	-	-	29.8	24.0	-	89.5
SC [30]	72.0	54.6	40.4	29.8	24.5	-	95.9
MSM [54]	73.4	56.7	43.0	32.6	25.4	54.0	100.2
G-LSTM [26]	67.0	49.1	35.8	26.4	22.7	-	-
Adap [13]	74.2	58.0	43.9	33.2	26.6	-	108.5
*Att2in** [33]	-	-	-	31.3	26.0	54.3	101.3
SCA [55]	71.9	54.8	41.1	31.1	25.0	53.1	95.2
UD-Base* [56]	74.5	-	-	33.4	26.1	54.4	105.4
Convcap [35]	71.1	53.8	39.4	28.7	24.4	52.2	91.2
AED-AR [57]	74.0	57.6	44.0	33.5	26.1	54.6	103.4
WICG [58]	67.9	50.2	36.7	27.1	22.6	49.9	81.8
HCVSA [59]	76.2	60.1	45.1	35.0	27.0	-	-
Baseline	72.8	56.1	42.5	32.3	25.1	53.2	101.6
Our_A	74.0	57.8	44.5	34.3	26.6	55.1	104.7
Our_R	73.9	58.7	44.4	34.0	26.7	54.7	105.8
Our_R_L	74.5	59.2	45.1	34.6	26.9	55.2	106.9
Our_A_R	75.1	60.0	46.0	35.2	27.5	**56.5**	109.4
Our_A_R_L	**75.9**	**60.3**	**46.5**	**35.8**	**27.8**	56.4	**111.3**

Table 2
The performance comparison with seven state-of-the-art methods on the Flickr30k dataset. Similarly, the results of ablated models (Baseline, Our_A, Our_R, Our_R_L and Our_A_R) and our full model (Our_A_R_L) are shown at the bottom of the table.

Method	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE-L	CIDEr
DeepVS [24]	57.3	36.9	24.0	15.7	15.3	-	24.7
Soft-Att [11]	66.7	43.4	28.8	19.1	18.5	-	-
Hard-Att [11]	66.9	43.9	29.6	19.9	18.5	-	-
ATT-FCN [31]	64.7	46.0	32.4	23.0	18.9	-	-
G-LSTM [26]	64.6	44.6	30.5	20.6	17.9	-	-
SCA [55]	66.2	46.8	32.5	22.3	19.5	-	-
Adap [13]	67.7	49.4	35.4	25.1	20.4	-	53.1
Baseline	66.8	48.7	34.9	24.5	19.3	45.2	51.3
Our_A	67.3	49.3	35.6	25.8	20.2	46.5	53.7
Our_R	68.0	50.7	36.2	26.5	20.7	47.0	55.3
Our_R_L	68.7	51.1	36.7	26.6	21.1	47.0	55.9
Our_A_R	69.2	51.3	37.5	27.3	21.3	48.2	56.3
Our_A_R_L	**69.8**	**51.7**	**37.8**	**27.7**	**21.5**	**48.5**	**57.4**

the subject word and object word mainly focus on the salient visual objects and backgrounds, respectively, while the nodes and their relationships corresponding to the relation word focus on both visual objects and backgrounds. That is to say, the relationships for the relation word indeed build the bridge between subject and object. For example, the relationships of green nodes selected to generate word standing focus on both the zebra and snow regions, and connect the subject word zebra and object word snow. To show the visualization differences between our model (with relationship learning, denoted as rel) and previous attention model (without relationship learning, denoted as att) in attention weight distribution, we also plot their distribution curves (with generated sentences) in Fig. 3. From the generated sentences, we can see that our model can produce more accurate semantic objects than the attention model (e.g., zebra-standing-snow vs zebra-standing). From the attention weight distribution, we can see that both our model and the attention model attend to some similar salient regions (e.g., the head region of zebra) when generating subject words (zebra and person). However, our model also attends to more different salient regions (e.g., the head region of zebra and the snow region) than the attention model when generated relation words (standing and riding). These results further indicate our model can boost previous attention model by attending to different context objects especially in generating relation words.

4.7. Performance on MSCOCO online testing server

To make a full comparison with other state-of-the-art methods, we have submitted Our_A_R (with visual representations) to the official MSCOCO evaluation server and obtain the model performance on the official testing set. Table 3 reports the performance leaderboard of published state-of-the-art methods and ours on the online MSCOCO test server. A test image on the leaderboard testing sets consists of five human-annotated captions (c5) or forty human-annotated captions (c40). In the experiment, Our_A_R does not use more complicated deep CNN models (i.e., ResNet-152 [45]) than the compared methods (e.g., MSM [54]). In addition, Our_A_R also does not use any reinforcement learning based objective function [33] which can apparently improve the performance of all evaluation metrics, and does not utilize model ensemble technique

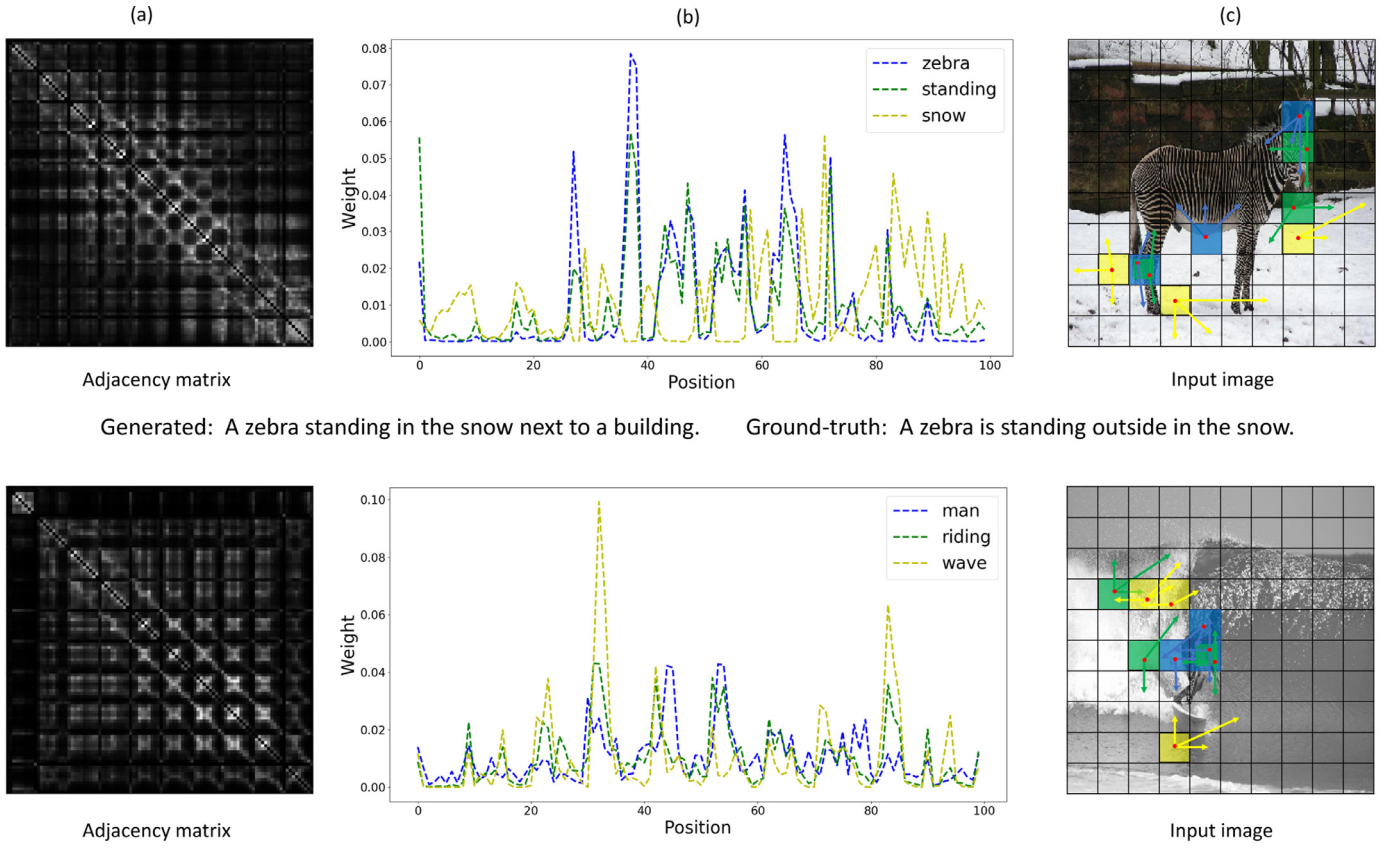

Fig. 2. Illustration of relationship adjacency matrix, attention weight distribution over 100 graph nodes when generating three subject-relation-object words and strong relationships corresponding to the mostly attended three graph nodes for two test images. The learned edge strengths/weights denote the probabilities of the relationships existing between any two graph nodes. The sentences generated by our proposed Our_A_R_L and human-annotated ground-truth sentences are also attached.

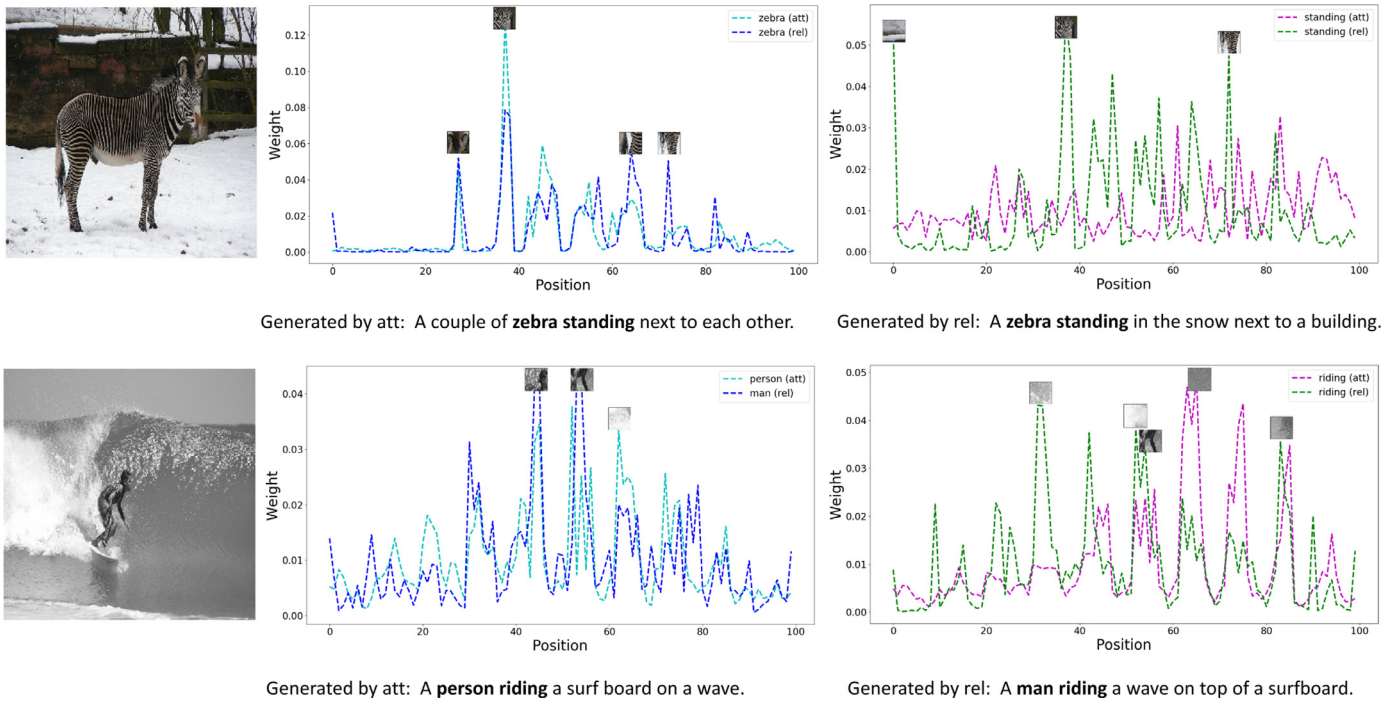

Fig. 3. Visualization differences between our model (with relationship learning, denoted as rel) and attention model (without relationship learning, denoted as att) in attention weight distribution when generating key words.

Table 3
The performance comparison with previous state-of-the-art image captioning methods on the online MSCOCO testing server. Here we directly cite most results from Lu et al. [13] to make a fair comparison.

Method	B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
NIC [8]	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946
MSCap [60]	0.543	0.819	0.407	0.710	0.308	0.601	0.248	0.339	0.526	0.680	0.931	0.937
mRNN [25]	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 [23]	0.548	0.804	0.409	0.695	0.306	0.585	0.247	0.335	0.528	0.678	0.921	0.934
HardA [11]	0.528	0.779	0.383	0.658	0.277	0.537	0.241	0.322	0.516	0.654	0.865	0.893
ATTF [31]	0.565	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958
ERD [12]	0.550	0.812	0.414	0.705	0.313	0.597	0.256	0.347	0.533	0.686	0.965	0.969
MSM [54]	0.575	0.842	0.436	0.740	0.330	0.632	0.256	0.350	0.542	0.700	0.984	1.003
Adap [13]	0.584	0.845	0.444	0.744	0.336	0.637	0.264	0.359	0.550	0.705	1.042	1.059
Ours	**0.589**	**0.856**	**0.450**	**0.756**	**0.343**	**0.647**	**0.270**	**0.364**	**0.555**	**0.710**	**1.061**	**1.064**

Table 4
The performance comparisons of our proposed Our_R with other variant methods on the MS COCO dataset in terms of BLEU@4 and CIDEr metrics.

Method	BLEU@4	CIDEr
Our_V(w/o R)	32.3%	101.6%
Our_R(w/o V)	34.0%	105.8%
Our_R(w/ V)	34.1%	105.7%
Our_R(GNN)	34.0%	105.8%
Our_R(GNN-4)	32.8%	103.0%
Our_R(GNN-8)	33.3%	104.6%
Our_R(FFCL)	32.6%	102.7%

(e.g., Adap [13]) to boost the performance. Compared with the 9 methods on the leaderboard, Our_A_R still outperforms them in terms of all evaluation metrics on both c5 and c40 testing sets.

4.8. Model analysis

Relationship representations. To investigate whether relationship representations can replace visual representations or not, we present the experimental results in the first three rows of Table 4. From these results, we can see that Our_R(w/o V) which employs only relationship representations achieves comparable results with Our_R(w/ V) which employs both relationship representations and visual representations in the evaluation metrics. Compared with Our_V(w/o R) which employs only visual representations, Our_R(w/o V) clearly outperform it by a large margin in the evaluation metrics. The comparison results indicate the learned relationship representations R can replace visual representations V for image captioning.

Other graph structures. To figure out how the other graph structures affect model performance, we present the experimental results of different graph structures in the last four rows of Table 4. Our_R(GNN) is implemented using graph neural network with fully-connected edges, Our_R(GNN-4) and Our_R(GNN-8) are implemented using graph neural network with 4-neighborhood and 8-neighborhood connections respectively, and Our_R(FFCL) is implemented using flatten and fully connected layers which is generally considered very close to Our_R(GNN). It can be seen that Our_R(GNN) achieves much better results than Our_R(GNN-4, Our_R(GNN-8) and Our_R(FFCL) in the evaluation metrics, which further proves the effectiveness of our graph model.

4.9. Analysis of parameters

To investigate the effect of two important parameters (iteration step T for the GNN and size of feature map for the attention model) for image captioning on the MS COCO dataset, we design several comparison experiments. Table 5 shows the results of our proposed Our_R model under different parameter settings. The first four rows of Table 5 lists the results of the iteration step T in the range of 3, 4, 5 and 6, and the last three rows of Table 5 shows the results of the size of feature map in the range of 6, 8 and 10. From these results, we can observe that increasing the iteration step T and the size of feature map can lead to performance improvements. In particular, the performance of our proposed Our_R does not increase too much when iteration step T is increased to a level. However, the number of parameters increases exponentially when the iteration step T and the size of feature map are increased. To make a tradeoff between performance and model complexity, we empirically set the iteration step T and the size of feature map to 4 and 10 in our experiments, respectively.

Table 5
The performance comparisons of our proposed Our_R under different parameter settings on the MS COCO dataset in terms of BLEU@4 and CIDEr metrics. Note that the number * in "T-*" and "F-*" denotes the number of the iteration step T for the GNN or the size of feature map for the attention model respectively. The first four rows show the results of the iteration step T with different parameter values (3, 4, 5 and 6), and the last three rows show the results of the size of feature map with different parameter values (6, 8 and 10).

Method	BLEU@4	CIDEr
Our_R(T-3)	33.5%	104.7%
Our_R(T-4)	34.1%	105.9%
Our_R(T-5)	34.2%	105.7%
Our_R(T-6)	34.2%	105.9%
Our_R(F-6)	32.7%	103.3%
Our_R(F-8)	33.4%	104.6%
Our_R(F-10)	34.1%	105.9%

5. Conclusions

In this paper, we have presented an image captioning method which consists of two novel components: graph based visual relationship modelling and context-aware attention mechanism. The visual relationship modelling is implemented via a graph neural network which recurrently passes the messages from adjacent nodes across time. The context-aware attention mechanism is implemented by a LSTM to memorize its previously attended visual information. Compared with the state-of-the-art methods, our proposed method can attend to both specific visual objects in an image and the implicit visual relationship among the visual objects of an image and take into account what has been previously attended to. We have evaluated the effectiveness of our proposed method on two public benchmark datasets: MS COCO and Flickr30K. In the experiments, our proposed method consistently outperforms the state-of-the-art methods in terms of all evaluation metrics on both datasets. We further visualize the spatial attention maps and gen-

erated sentences to better understand our method, which indicates that our proposed method learns information consistent with human perception. In the future, we will aim to integrate explicit visual relationship into our method. Furthermore, our proposed method can also be applied to other vision-to-language tasks such as visual question answering and visual dialogue.

Declaration of Competing Interest

None.

Acknowlgedgments

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61420106015, 61572504), and Australian Research Council (ARC) Grants (DP160103675).

References

[1] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, W. Dong, Image caption generation with part of speech guidance, Pattern Recognit. Lett. 001 (2017) 1–9, doi:10.1016/j.patrec.2017.10.018.
[2] P. Kinghorn, L. Zhang, L. Shao, A hierarchical and regional deep learning architecture for image description generation, Pattern Recognit. Lett. 003 (2017) 1–9, doi:10.1016/j.patrec.2017.09.013.
[3] S. Aditya, Y. Yang, C. Baral, Y. Aloimonos, C. Fermüller, Image understanding using vision and reasoning through scene description graph, Comput. Vision Image Understanding (2017), doi:10.1016/j.cviu.2017.12.004.
[4] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[5] X. Bai, B. Shi, C. Zhang, X. Cai, L. Qi, Text/non-text image classification in the wild with convolutional neural networks, Pattern Recognit. 66 (2017) 437–446.
[6] H. Choi, K. Cho, Y. Bengio, Context-dependent word representation for neural machine translation, Comput. Speech Lang. 45 (2017) 149–160.
[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1724–1734.
[8] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge, IEEE Trans. Pattern Anal. Mach.Intell. 39 (4) (2017) 652–663.
[9] K. Fu, J. Jin, R. Cui, F. Sha, C. Zhang, Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts, IEEE Trans. Pattern Anal. Mach.Intell. 39 (12) (2017) 2321–2334.
[10] L. Li, S. Tang, Y. Zhang, L. Deng, Q. Tian, GLA: Global-Local attention for image description, IEEE Trans. Multimed. 20 (3) (2018) 726–737.
[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 2048–2057.
[12] Z. Yang, Y. Yuan, Y. Wu, W.W. Cohen, R.R. Salakhutdinov, Review networks for caption generation, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 2361–2369.
[13] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3242–3250.
[14] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2009) 61–80.
[15] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 15–29.
[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Babytalk: Understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach.Intell. 35 (12) (2013) 2891–2903.
[17] Y. Yang, C.L. Teo, H. Daumé III, Y. Aloimonos, Corpus-guided sentence generation of natural images, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454.
[18] A. Gupta, Y. Verma, C.V. Jawahar, Choosing linguistics over vision to describe images, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 606–612.
[19] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, M. Mitchell, Language models for image captioning: The quirks and what works, in: Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics, 2015, pp. 100–105.
[20] V. Ordonez, G. Kulkarni, T.L. Berg, Im2Text: Describing images using 1 million captioned photographs, in: Proceedings of the Advances in Neural Information Processing Systems, 2011, pp. 1143–1151.
[21] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi, Collective generation of natural image descriptions, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, pp. 359–368.
[22] P. Kuznetsova, V. Ordonez, T. Berg, Y. Choi, Treetalk: Composition and compression of trees for image descriptions, Trans. Assoc. Comput.Ling. 2 (1) (2014) 351–362.
[23] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach.Intell. 39 (4) (2017) 677–691.
[24] K. Andrej, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, IEEE Trans. Pattern Anal. Mach.Intell. 39 (4) (2017) 664–676.
[25] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks (m-RNN), in: Proceedings of the International Conference on Learning Representations, 2015.
[26] X. Jia, E. Gavves, B. Fernando, T. Tuytelaars, Guiding the long-short term memory model for image caption generation, in: Proceedings of the International Conference on Computer Vision, 2015, pp. 2407–2415.
[27] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the International Conference on Learning Representations, 2015.
[28] P. Rodríguez, G. Cucurull, J.M. Gonfaus, F.X. Roca, J. Gonzalez, Age and gender recognition in the wild with deep attention, Pattern Recognit. 72 (2017) 563–571.
[29] E. Fidalgo, E. Alegre, V. González-Castro, L. Fernández-Robles, Boosting image classification through semantic attention filtering strategies, Pattern Recognit. Lett. 112 (2018) 176–183.
[30] L. Zhou, C. Xu, P. Koch, J.J. Corso, Watch what you just said: Image captioning with text-conditional attention, in: Proceedings of the ACM International Conference on Multimedia Workshops, 2017, pp. 305–313.
[31] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659.
[32] Q. Wu, C. Shen, L. Liu, A. Dick, A. van den Hengel, What value do explicit high level concepts have in vision to language problems? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 203–212.
[33] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1179–1195.
[34] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, K. Murphy, Improved image captioning via policy gradient optimization of SPIDEr, in: Proceedings of the International Conference on Computer Vision, 2017, pp. 873–881.
[35] J. Aneja, A. Deshpande, A.G. Schwing, Convolutional image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5561–5570.
[36] Q. Wang, A.B. Chan, CNN + CNN: convolutional decoders for image captioning, CoRR (2018). arXiv: abs/1805.09019.
[37] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang, Scene graph generation from objects, phrases and region captions, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1261–1270.
[38] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 684–699.
[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vision 123 (1) (2017) 32–73.
[40] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated graph sequence neural networks, in: Proceedings of the International Conference on Learning Representations, 2016.
[41] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the International Conference on Learning Representations, 2017.
[42] R. Li, S. Wang, F. Zhu, J. Huang, Adaptive graph convolutional neural networks, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 3546–3553.
[43] X Wang, A. Gupta, Videos as space-time region graphs, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 399–417.
[44] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: Proceedings of the International Conference on Learning Representations, 2018.
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[46] V.G. Satorras, J.B. Estrach, Few-shot learning with graph neural networks, in: Proceedings of the International Conference on Learning Representations, 2018.
[47] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, CoRR (2014). arXiv: abs/1409.2329.
[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.
[49] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput.Ling. 2 (2014) 67–78.

[50] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.
[51] K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318.
[52] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the 9th Workshop on Statistical Machine Translation, 2014, pp. 376–380.
[53] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Proceedings of the Association for Computational Linguistics, 2004, pp. 74–81.
[54] T. Yao, Y. Pan, Y. Li, Z. Qiu, T. Mei, Boosting image captioning with attributes, in: Proceedings of the International Conference on Computer Vision, 2017, pp. 4904–4912.
[55] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6298–6306.
[56] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
[57] X. Chen, L. Ma, W. Jiang, J. Yao, W. Liu, Regularizing RNNs for caption generation by reconstructing the past with the present, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7995–8003.
[58] M. Tanti, A. Gatt, K.P. Camilleri, Where to put the image in an image caption generator, Nat. Lang. Eng. 24 (3) (2018) 467–489.
[59] M. Khademi, O. Schulte, Image caption generation with hierarchical contextual visual spatial attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1943–1951.
[60] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J.C. Platt, C. Lawrence Zitnick, G. Zweig, From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

Junbo Wang received the B.S. degree from software engineering of Northeastern University in 2014. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include video summarization, video captioning, image captioning and deep learning.

Wei Wang received the B.E. degree in Department of Automation from Wuhan University in 2005, and Ph.D. degree in School of Information Science and Engineering at the Graduate University of Chinese Academy of Sciences (GUCAS) in 2011. Since July 2011, Dr. Wang has joined NLPR as an assistant professor. His research interests focus on computer vision, pattern recognition and machine learning, particularly on the computational modeling of visual attention, deep learning and multimodal data analysis. He has published more than ten papers in the leading international conferences such as CVPR and ICCV.

Liang Wang (SM'09) received both the B.S. and M.S. degrees from Anhui University in 1997 and 2000 respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he has been working as a Research Assistant at Imperial College London, United Kingdom and Monash University, Australia, a Research Fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full Professor of Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition and computer vision. He has widely published at highly-ranked international journals such as IEEE TPAMI and IEEE TIP, and leading international conferences such as CVPR, ICCV and ICDM. He is an associate editor of IEEE Transactions on SMC-B. He is currently an IAPR Fellow and Senior Member of IEEE.

Zhiyong Wang received his B. Eng. and M. Eng. Degrees in electronic engineering from South China University of Technology, Guangzhou, China, and his Ph.D. degree from Hong Kong Polytechnic University, Hong Kong. He is currently an Associate Professor and Associate Director of the Multimedia Laboratory with the School of Information Technologies, The University of Sydney, Sydney, Australia. His research interests focus on multimedia computing, including multimedia information processing, retrieval and management, Internet-based multimedia data mining, human-centered multimedia computing, and pattern recognition.

David Dagan Feng received his M. Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, CA, USA, in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. He is Director of the Biomedical & Multimedia Information Technology Research Group, and Research Director of the Institute of Biomedical Engineering and Technology at the University of Sydney, Sydney, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He has served as Chair of the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, has organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions. He is a Fellow of IEEE and Australian Academy of Technological Sciences and Engineering.

Tieniu Tan received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He is currently a professor in the Center for Research on Intelligent Perception and Computing, NLPR, CASIA, China. He has published more than 450 research papers in refereed international journals and conferences in the areas of image processing, computer vision and pattern recognition, and has authored or edited 11 books. His research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a fellow of the CAS, the TWAS, the BAS, the IEEE, the IAPR, the UK Royal Academy of Engineering, and the Past President of IEEE Biometrics Council.