Visual Relational Reasoning for Image Caption

Haolei Pei
School of Information and Technology
Zhejiang Sci-Tech University
HangZhou 310018, China
haleypei@outlook.com

Qiaohong Chen*
School of Information and Technology
Zhejiang Sci-Tech University
HangZhou 310018, China
chen_lisa@zstu.edu.cn

Ji Wang
School of Information and Technology
Zhejiang Sci-Tech University
HangZhou 310018, China
jiwang2014@outlook.com

Qi Sun
School of Information and Technology
Zhejiang Sci-Tech University
HangZhou 310018, China
sunqi@vip.sina.com

Yubo Jia
School of Information and Technology
Zhejiang Sci-Tech University
HangZhou 310018, China
jiayubo1964@163.com

***Abstract*—Recently, various attention-based networks have achieved state-of-art results on image captioning tasks. However, this simple mechanism is insufficient to modelling and reasoning the relationships between the visual regions required for scene understanding. In this research, we propose a visual relational reasoning module to implicit learning semantic and spatial relationships between pairs of relevant visual objects and infers the feature output that is most relevant to the currently generated word. Furthermore, a context gate is introduced to dynamically control the contribution of visual region attention modules and visual relational reasoning module which allows predicting different words according to different type of features (visual or visual relationship). We evaluate our model on the MSCOCO dataset and achieved state-of-the-art results. Qualitative analysis shows that our visual relational reasoning model can dynamically model and reason the most relevant features of different types of generated words and improve the quality of the caption.**

I. Introduction

As a high-level visual task, the image caption which aims to describe the correctly content of an image has received more and more attention from academia. This technology has a great potential impact on a wide range of applications such as helping visually impaired or robot visual understanding. Image caption is a challenging task for machines because it not only requires a comprehensive understanding of objects, scene, and their mutual relations, but also needs to describe the content of an image with semantically and syntactically correct sentences.

Recently, the encoder-decoder based image caption model with attention mechanism has achieved extraordinary performances. The convolutional neural network (CNN) is often used as an encoder to transform a image into a feature vector, and then input it into the Long Short-Term Memory (LSTM) Network decoder to generate a word at each time step. Based on the Soft attention mechanism [1], [2] proposed a visual sentinel mechanism, which allows the model to choose whether to focus on the image at each time step. [3] proposes a model that combines bottom-up and top-down attention model, which can attend to images at the level of salient objects. [4] proposes a multi-stage prediction framework with increasingly refined attention weights for image captioning. [5] proposes a Hierarchical Attention Network that calculate the attention of different modal visual features and perform feature fusion through parallel multivariate residual modules.

However, the focus of these models is still to make better use of the visual objects feature extract from the images, ignore the impact of the relationship among visual objects during generation. For instance, when the model generates a description "A young girl holding a tennis racquet on a tennis court", it should figure out the relationship "holding" between "girl" and "racquet", and the relationship "on" between "girl" and "tennis court". On one hand, we think that in a human system, when a person is describing a picture, it is usually necessary to consider the relationship between the pair of objects and the overall context to accurately describe the relationship. On the other hand, this reasoning process is not needed when describing specific objects. Therefore, we think that describing different types of words should be handled by different modules, which requires a mechanism to control.

To resolve the above restrictions, we propose a visual relational reasoning model, which is inspired by the relational network proposed in [6]. First, to improve the efficiency, we use the visual attention module to select the visual regions that are most related to the currently generated words. Then, Our relational reasoning model can simultaneously encode the semantic, spatial, and contextual relationships between these object pairs. Subsequently, we use an attention mechanism to dynamically reasoning the visual relationship features that select related ones for the LSTM decoder. Finally, in order to cooperate with the attention mechanism that is good at generating visual object words, a context gating mechanism is introduced to dynamically control the contributions of different types of features. It can make gradients of different types of words back propagate correctly to different modules, so that the model can be trained end-to-end. Our visual relational reasoning model is designed based on the R-CNN-LSTM

*Corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

framework, which combines the bottom-up regional attention model in [3]. Faster R-CNN as a image encoder can generate visual region feature and their spatial feature. The visual object features output by the visual attention module and the visual relationship features output by the relational reasoning module will be input in context gate to decide which feature to use to generate the next word.

Overall, the main contributions of this paper are summarized as follows:

- We propose a visual relational reasoning model that can implicitly model the semantic, spatial, and contextual relationships between visual regions of related, and infer a relationship feature that is most relevant to the current generated word.
- We introduce a context gate mechanism to adaptively control the contribution of features of different types.
- We conduct a number of experiments on the MSCOCO dataset and the results show that our model outperforms the state-of-the-art approaches.

II. Related Work

Recent image captioning work is mainly based on encoder-decoder framework [1], [7]–[11] that uses CNN to encode image features and then enter RNN to generate sentences. Specifically, [8] first proposed an end-to-end image caption network based on the seq-to-seq model framework, and utilizing LSTM as a decoder to generate sentences. Based on [8], [1] further proposes soft and hard attention mechanism, which allows the model to focus on the relevant regions of the image at each generation time step. Because there are non-visual words in the sentence that do not need attention, [2] proposed a visual sentinel mechanism, which allows the model to choose whether to focus on the image at each time step. Recently, [12] extracts semantic attributes from images and generates semantic attention [11] at each generation time step to enhance the quality of caption. Most recently, [3] proposes a model that combines bottom-up and top-down attention mechanism, which can attend to images at the level of salient objects. [4] proposes a multi-stage prediction framework with increasingly refined attention weights for image captioning. [5] proposes a Hierarchical Attention Network that calculate the attention of different modal visual features and perform feature fusion through parallel multivariate residual modules. However, these models ignore the visual relationship between regions. The spatial information of these visual regions and the relationship between them are also neglected.

[13] first proposed work to exploring the visual relationships between image regions. It uses graph convolutional networks and LSTM architectures for the first time. This structure uses GCN to build semantic and spatial graphs between image regions, and then inputs the extracted semantic and spatial relationship features into an LSTM-based decoding architecture with attention mechanism to generate sentences. Its spatial and semantic graph edges are all generated by pre-trained networks using the Visual Genome dataset [14]. The 11 categories in spatial relations and 21 categories of semantic relations are both predefined. However, this model has some limitations. First, its semantic and spatial relationship graphs need to be trained separately. Secondly, the relational features of this model are mainly generated by pre-trained models, GCN is only a feature refinement model that connects pre-trained networks and image caption networks. This main network of this model has no ability to dynamically explore the relationships between image regions during the training process.

[6] recently proposed the Relation Network (RN) architecture for visual question answering, designed specifically for augmenting relational reasoning performance. The RN reasons about all image regions pairs explicitly. Image region embeddings are generated with a CNN, and questions are embedded with an LSTM. For each region pair, the embeddings of the region pairs and the question embedding are concatenated, and passed through a MLP, generating a feature vector for that region pair. These vectors are then summed to a final embedding used for classification.

Inspired by RN's encoding mechanism of relationship features, we design a visual relational reasoning model base on object pair that dynamically encode and infer the implicitly semantic, spatial and contextual relationships between visual regions. Compare with using GCN mechanism to encoding multiple relationship graphs for all visual regions, we have designed a rank embedding mechanism that uses a common attention mechanism to generate a part of the visual regions that are most relevant to the current generated word. The rank embedding mechanisms can greatly reduce the computational complexity of the model. At the same time, the model can simultaneously encode implicitly semantic, spatial and contextual relationships, which does not need predefined relationship classes. Finally, with the context gating mechanism, the relational reasoning model can be used with other attention models to help generate the different types of words.

In summary, our method has two advantages over previous models. First, we design a relational reasoning network that using the visual region's visual features and spatial features can dynamically generate the relationship features, and then reason the most relevant output through the attention mechanism. Second, we introduced the context gating mechanism to adaptively control the contribution of features on different types, so this allows the relational reasoning module to cooperate with the visual attention module to generate high-quality captions.

III. Model

A. Over All

Given an image I, the image captioning model needs to generate a caption sequence $S = \{w_1, w_2, \ldots, w_T\}, w_t \in D$, where D is the vocabulary dictionary and T is the sequence length. As illustrated in Fig. 1, We adopt the R-CNN-LSTM proposed by [3] for image captioning. In particular, we use an object detection module (Faster R-CNN) to detect objects $V = \{v_1, v_2, \ldots, v_N\}$, each $v_n \in R^{D_v}$ is a d-dimensional visual object vector, and its coordinates $b = (x, y, w, h)$ of the

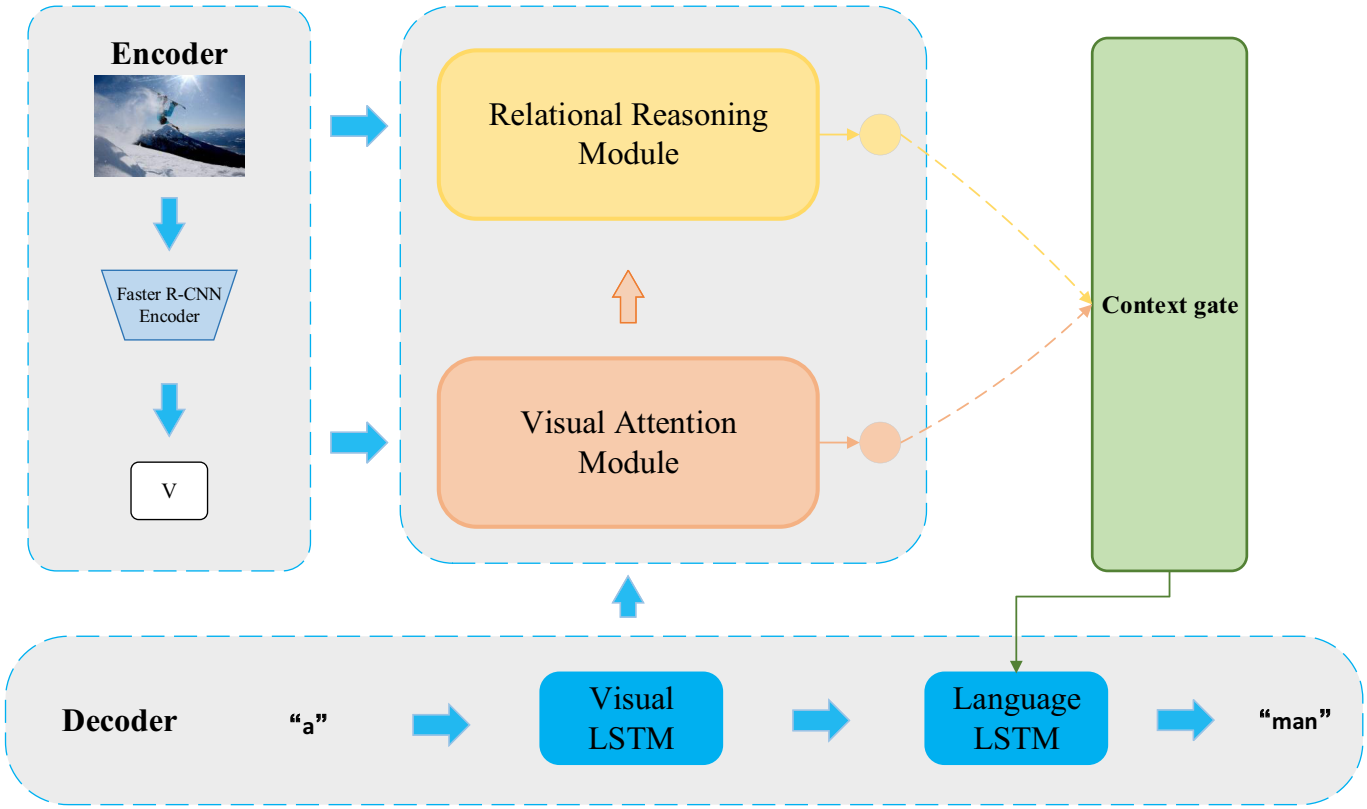


Fig. 1. The overview framework of our relational reasoning model, which is composed of Encoder module, Visual Attention module, Relational Reasoning module, Context gate and Decoder module. The model takes the images and the words generated at last time step as input and outputs the next words.

bounding box with center (x, y), width w and height h within image. As a semantic decoder, RNN is leveraged to guide the generation of attention and caption sequences. based on the top-down attention framework in [3], we adopt a 2-layer LSTM [15] as a decoder in this paper.

To decouple attention guidance and sequence generation, We design the Visual Attention Module and the Relation Reasoning Module to allow them to process image features in parallel. The visual attention module is applied to generate attention features in every time step, while the relational reasoning module aims to encode and infers the visual relationship among the visual object in an image.

We construct a cascaded LSTM structure that includes a visual LSTM and a language LSTM. The visual LSTM is applied to perceive global information of images and guide visual attention module to generate attention features and guide relational reasoning module to obtain the visual relationship features. While the language LSTM guides caption generation. Finally, we construct a context gate that control the contribution of object context and relation context. The overall structure of our model is shown in Fig. 1. During the generation process, the visual encoder extracts visual object features. The visual LSTM reviews the global information of the image at each moment and guides attention models and relational reasoning module to refine features. The features of the different modules are entered into the context gate to control the features to be output. The language LSTM generates a word at each moment given last word and multimodal features.

The process can be defined by the following formulas:

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

$$h_t^V = LSTM_V([h_{t-1}^L, \bar{v}, E(W_t)]) \tag{2}$$

$$A_t, \alpha_t = Attention(h_t^V, V) \tag{3}$$

$$R_t = Relation(h_t^V, V, b, \alpha_t) \tag{4}$$

$$C = Contextgate(h_t^V, R_t, A_t, b) \tag{5}$$

$$h_t^L = LSTM_L([C, h_t^V]) \tag{6}$$

$$W_t = \arg\max_S softmax(W_o h_t^L + b_o) \tag{7}$$

where $Encoder(I)$ represents feature extractor and $E()$ is the embedding function which maps the one-hot representation into the embedding space. $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$ is the mean-pooled region feature. $LSTM_V$, $Attention()$, $Relation()$, $Contextgate()$ and $LSTM_L$ represent visual LSTM, visual attention module, relational reasoning module, context gate and language LSTM, respectively.

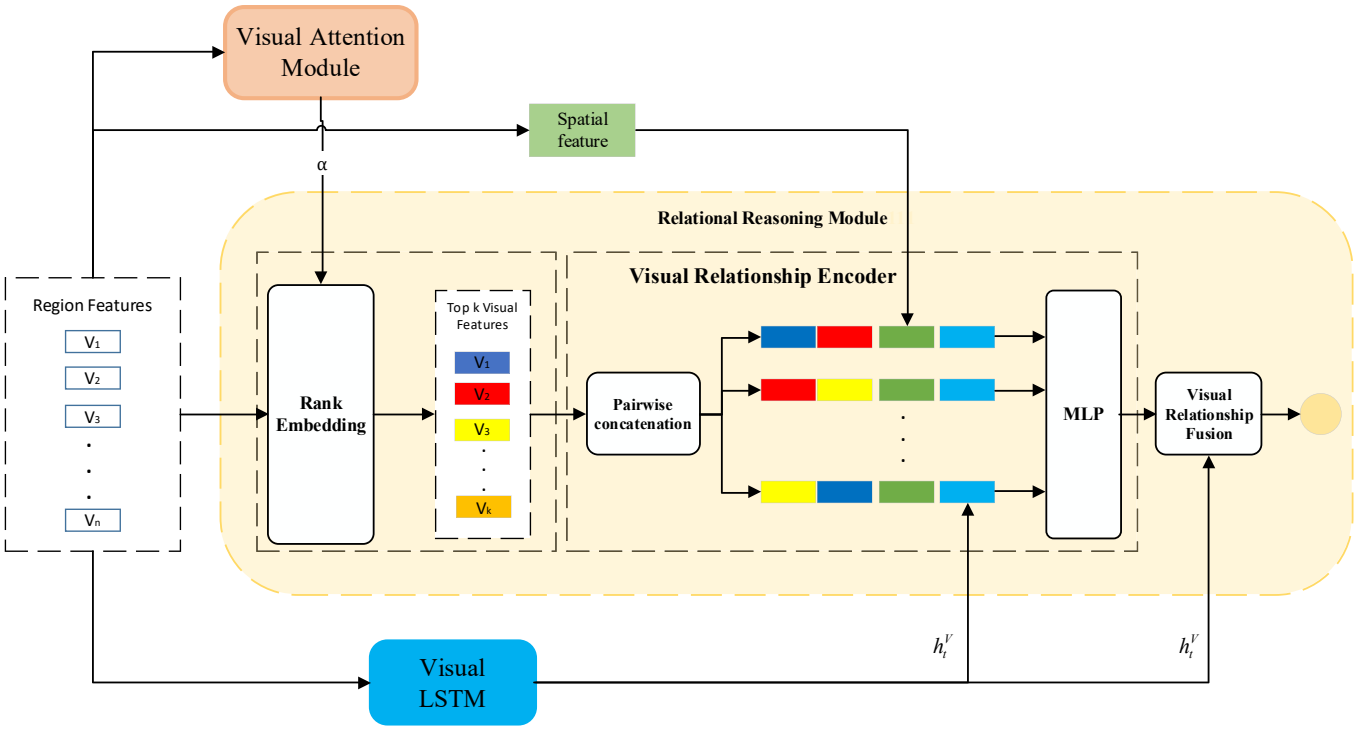


Fig. 2. The illustration of Relational Reasoning Module. First the rank embedding mechanism will select the highest K regions of the attention value obtained by the visual attention module. Then the visual relationship feature encoder can build pairwise combinations of k region proposals, where in turn we get K(K-1) possible region pairs. Finally, the visual relationship fusion can dynamically reasoning the visual relationship features that select related ones for the LSTM decoder.

B. Visual Attention Module

The soft attention mechanism [1] is introduced into our framework as visual attention module. The traditional attention mechanism has fully demonstrated its advantages in generating visual words. Therefore, the purpose of the visual attention module is to focus on the visual region features that are most relevant to the visual word at the current time step, rather than considering the relationship among each visual region features. Given the visual region features V and the output h_t^V of the visual LSTM, our visual attention module uses the following formula to normalize attention weights:

$$a_t = W_a^t tanh(W_{va}V + W_{ha}h_t^V) \tag{8}$$

$$\alpha_t = softmax(a_t) \tag{9}$$

where $W_{va} \in \mathbb{R}^{H\times V}$, $W_{ha} \in \mathbb{R}^{H\times M}$and $W_a \in \mathbb{R}^H$are learned parameters. The attended image feature used as input to the language LSTM is calculated as a weighted sum of all input feature:

$$A_t = \sum_{i=1}^{K} \alpha_t v_i \tag{10}$$

C. Relational Reasoning Module

We extend the model with a relational reasoning module, based on the Relation Network (RN) architecture [6]. Intuitively, this module can dynamically encode and reasoning the visual relationship at each time step t, by explicitly considering combinations of image region pairs. The visual relational reasoning module consists of three parts: rank embedding, visual relationship encodier, and visual relationship fusion. (1) The rank embedding mechanism can use visual attention module to select a subset of visual region features. (2) The visual relationship encoder can encode the semantic, spatial and context relationships between object pairs. (3) The visual relationship fusion can infer the visual relationship features that are most relevant to the visual relationship words at the current time step. The overall relational reasoning module framework is illustrated in Fig. 2.

1) Rank Embedding: The Rank Embedding mechanism aims to augment the relationship encoding efficient of the follow-up model. This design is mainly used to combine the visual attention module to filter the visual region features, which can greatly reduce the computational complexity and parameters while improving the efficiency of the model. First, the attention weights α are used to select the m most relevant regions $V_a \in V$ where $|V_a| = k$,$k < n$, and the selected regions correspond to the k highest attention weights.

2) Visual Relationship Encoder: The Visual relationship encoder can encode the semantic, spatial and context relationships between the k visual region features selected in the previous step.

Specifically, illustrated by the middle part in Fig. 2, we first build pairwise combinations of visual features $\{v_1, v_2, \ldots, v_k\} \in V_a$, where in turn we get $k(k-1)$

possible visual object feature pairs $[v_i, v_j]$, set $i \neq j$, while $i, j \in [1, k]$,and concatenation together with hidden unit h_t^V and spatial feature . We define the pairwise object relationship encoder as a composite function below:

$$r_{i,j} = MLP([v_i, v_j, W_S s_{ij}, h_t^V]) \tag{11}$$

Where $W_S \in \mathbb{R}^{D_H \times 6}$is transformation matrices, and the spatial feature s_{ij} is defined similarly to [16] as:

$$s_{ij} = \left[\frac{x_i - x_j}{\sqrt{w_j h_j}}, \frac{y_i - y_j}{\sqrt{w_j h_j}}, \sqrt{\frac{w_i h_i}{w_j h_j}}, \frac{w_j}{h_j}, \frac{w_i}{h_i}, \frac{b_j \cap b_i}{b_j \cup b_i}\right] \in R^6 \tag{12}$$

3) Visual Relation Fusion: The visual relationship fusion is used to reasoning the related relationship features for output. We choose the attention mechanism to infer the relationship feature r. The formula is as follows:

$$z_t = W_k^R tanh(W_{va}^R r_k + W_{ha}^R h_t^V) \tag{13}$$

$$\alpha_t^R = softmax(z_t) \tag{14}$$

Where, $W_k^R \in \mathbb{R}^{1 \times D_a}, W_{va}^R \in \mathbb{R}^{D_a \times D_v}, W_{ha}^R \in \mathbb{R}^{D_a \times D_h}$are transformation matrices. W_{va}^R and W_{ha}^R map the visual relationship feature r_k and language feature ht^V into the same shared feature space. Based on the weight matrix, relational inference features R_t are obtained by weighted addition at each time step.

$$R_t = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_{t,i,j}^R r_{i,j} \tag{15}$$

D. Context Gating

Inspired by the gating mechanism in LSTM [15] and the work [17] in dense video captioning, we introduce a context gating mechanism into our model dynamically to control the contribution of visual region level context and visual relation level context on the word prediction. When obtaining the visual relationship features R and the visual attention features A, we learn a context gate to dynamically control them. First, we project the two different features into the same space:

$$\widetilde{C}_R = tanh(W_R R) \tag{16}$$

$$\widetilde{C}_V = tanh(W_A A) \tag{17}$$

where W_R, W_A are the transformation matrices. The context gate is then calculated by a nonlinear sigmoid function:

$$gctx = \sigma(W_g[\widetilde{C}_R, \widetilde{C}_V, h_t^V]) \tag{18}$$

where h_t^V is the previous visual LSTM state and gctx is a 2048-d weight vector. We could fuse the relation features and the visual object features as follows:

$$C = [(1 - gctx) \circ R, gctx \circ A] \tag{19}$$

TABLE I
THE PERFORMANCE OF THE ABLATION EXPERIMENT ON RELATION AND OBJECT FEATURES WITH DIFFERENT COMBINATION.

Model	B-1	B-4	M	R	C	S
Add	75.8	35.2	27.1	56.2	111.3	20.1
Concat	76.3	35.7	27.4	56.4	113.5	20.2
Context gating	77.6	36.9	27.9	56.8	115.5	20.8

TABLE II
THE PERFORMANCE OF THE ABLATION EXPERIMENT ON DIFFERENT K COMBINATIONS IN RANK EMBEDDING MODULE.

m	B-1	B-4	M	R	C	S
3	76.4	35.3	27.2	56.3	111.2	20.1
5	76.3	35.7	27.4	56.4	113.5	20.3
7	77.6	36.9	27.9	56.8	115.5	20.8
11	77.3	36.4	27.7	56.7	115.4	20.6

E. Model Learning

Firstly, we adopt the usual cross entropy loss (XE) to optimize our model. Considering the XE is not the final metric for image caption task, we further adopt the CIDEr [18] as the objective function to finetune our model. Specially, we minimize the negative expectation score of CIDEr as follows:

$$L(\theta) = -E_{w^s \sim p_\theta}[CIDEr(w^s)] \tag{20}$$

According to [18], the expected gradient for single sample $w^s \sim p_\theta$ is:

$$\nabla_\theta L(\theta) \approx -(CIDEr(w^s) - CIDEr(w)) \nabla_\theta \log p_\theta(w^s) \tag{21}$$

where $w^s = (w_1^s, \ldots, w_T^s)$, w^s is the sequence sample from the model, $CIDEr(w)$ is the reward score obtained by predicted sequence.

IV. EXPERIMENTS

A. Datasets and Evaluation metrics

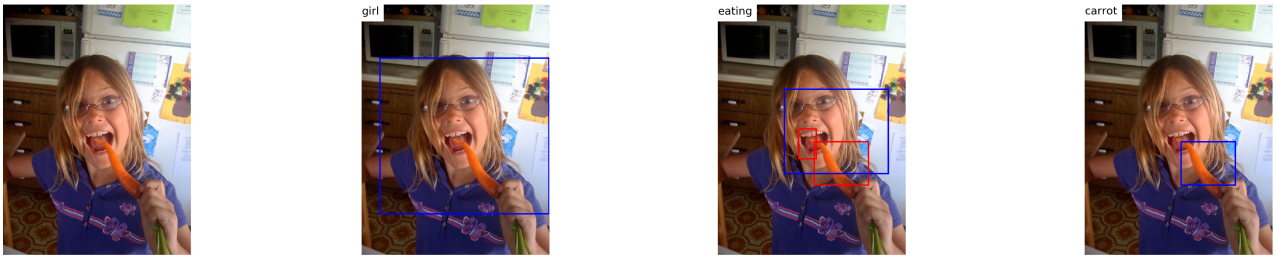
The MSCOCO dataset [19] is the largest public dataset for image caption, so we introduce it as a benchmark dataset. The dataset contains 123,287 images and is split into 82,783 for training and 40,504 for validation. Each image has 5 description of human annotations. For evalution, we use the Karpathy's splits [1] which contain 113,287, 5,000, and5,000 images for training, validation and evalution. To fairly evaluate the quality of the generated caption, we introduce the evaluation criteria widely applied in previous works: BLEU [20], METEOR [21], ROUGEL [22], CIDEr [18] and SPICE [23].

B. Image Features

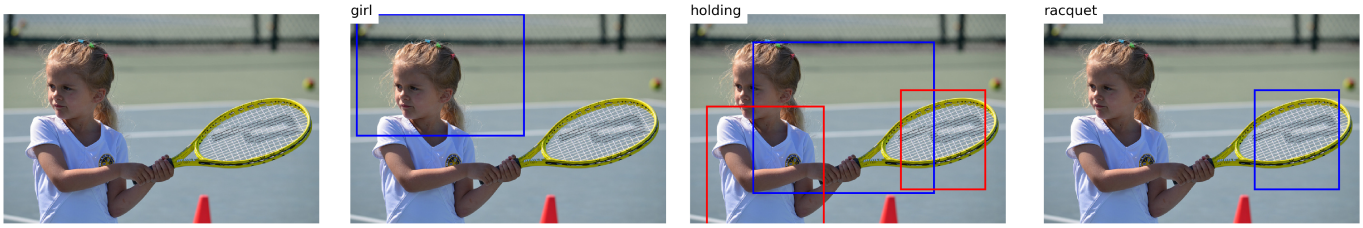
We use recent bottom up features [3] to represent our image as a region feature set. It generated by faster R-CNN [24] pretrained on Visual Genome [14] in conjunction with ResNet-101 [25]. First, Faster R-CNN detects top 36 highest confidence salient object regions in the image and generates corresponding bounding boxes. Then ResNet-101 is used to extract the 2048-dimensional region features in the feature map of the last convolution layer. bounding box coordinates are further used to calculate spatial features.

TABLE III
THE PERFORMANCE OF OUR MODEL ON THE MSCOCO KARPATHY'S TEST SPLIT.

Model	B-1	B-4	METEOR	ROUGE-L	CIDEr	SPICE
Att2in	-	31.3	26	54.3	101.3	-
Adaptive	74.2	33.2	26.6	-	108.5	19.5
NBT	75.5	34.7	27.1	-	107.2	20.1
Updown	77.2	36.2	27.0	56.4	113.5	20.3
Ours:VREA(XE)	**77.6**	**36.9**	**27.9**	**56.8**	**115.5**	**20.8**
Att2in	-	33.3	26.3	55.3	111.4	-
Updown	79.8	36.3	27.7	56.9	120.1	21.4
Ours:VREA(CIDEr)	**80.2**	**37.4**	**28.1**	**57.2**	**122.1**	**21.9**



a girl with glasses is eating a carrot.



a young girl holding a tennis racquet on a tennis court.

Fig. 3. We visualized the image regions with the highest attention weight in the visual attention module and the visual relational reasoning module for the relationship words. For the visual relationship words and visual words in the description, the blue box indicates the image region of the maximum attention weight generated by the visual attention module, and the red box indicates the image region pair of the maximum attention weight generated by the visual relational reasoning module.

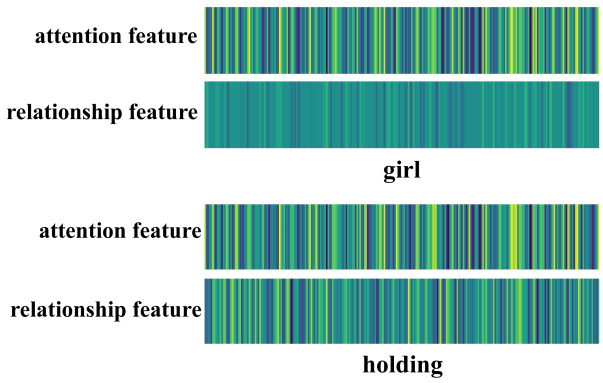


Fig. 4. The visualization of the gate values of the different types of words in context gate. The green and yellow colours represent the highest and lowest score respectively.

C. Implement Details

We first convert all the captions to low case and replace words less than 5 times with an UNKNOW token. Then, we build a word vocabulary with 10,010 unique words. The word embedding size is 1024. We set the hidden size of the two LSTM are both 1024, $k = 7$ objects for the relational reasoning module. The hidden size of the visual attention module is set to 512. Our whole model is trained by Adam [26] optimizer with batch size 128. For the training with cross entropy, we initially set the leaning rate as 5×10^{-4} and is reduced by 20% every 3 epochs. The maximum iteration is set as 40 epochs. When cross entropy training is over, we start the self-critical training and achieves best CIDEr score on validation set. We start the training with learning rate 5×10^{-5} reduced by 20% every 3 epochs for 30 epochs. At inference, beam search strategy is adopted and we set the beam size as 5. The training takes about 6 days on Nvidia Quadro M4000 GPU.

D. Quantitative Analysis

To better understand the effect of our context gating strategy, we carry out an ablation study and the results are exhibited in TABLE. I. Compared with the previous methods that apply concatenation or addition to integrate features, our context gate achieves an improvement of 3.6% and 1.7% in terms of the CIDEr metric. This indicates that our context gate can improve the quality of descriptions by dynamically control the contribution of different features.

To illustrate the influence of different parameter k in rank embedding module, we further carry out an ablation study and the results are shown in TABLE. II. We find that the number of regions m have a greater impact on performance, where k = 3 model is 1.6% lower than the CIDEr metric of k = 5 models. However, too large k does not necessarily improve the overall performance, so we choose k = 7 for our relational reasoning model.

For evaluation on MSCOCO dataset, we report the performance of our model in comparison with the current state-of-the-art methods: Adaptive [2], Att2in [27], NBT [28] and Updown [3]. TABLE. III demonstrate the results on the MSCOCO Karparthy test split. From the table, we can find that our model has advantages in all metrics. With XE objective, our model has an superiority over updown model with an improvement of 1.6% in terms of the CIDEr metric. With CIDEr objective,We can see that all metrics have improved by 1%-6%. Compared to the other models, our model has advantages over updown model with an improvement of 1.7% in terms of the CIDEr metric.

E. Qualitative Analysis

To qualitatively evaluate our proposed visual relational reasoning module, we visualized the image regions with the highest attention weight in the visual attention module and the visual relational reasoning module for the relationship words in generated descriptions in Fig. 3. For the visual relationship words and visual words in the description, the blue box indicates the image region of the maximum attention weight generated by the visual attention module, and the red box indicates the image region pair of the maximum attention weight generated by the visual relationship reasoning. In the word "eating" in the top of Fig. 3, the red region pair generated by the visual reasoning module correctly noticed the area mouth and carrot, while the blue area generated by the visual attention module incorrectly noticed the other areas. But the blue box generated by visual attention module can accurately notice the position of the visual words "girl" and "racquet" in the picture. In the word "holding" in the bottom of Fig. 3, the red region pair generated by the visual reasoning module correctly noticed the area arms and turnips, while the blue area generated by the visual attention module incorrectly noticed the other areas. But the blue box generated by visual attention module can accurately notice the position of the visual words "girl" and "carrot" in the picture. The above visualization can prove that our visual relational module can accurately encode and reason about relationships to better generate descriptions.

In order to qualitatively analyze our context gating mechanism, we exhibit the visualizations of the context gate when predicting the words "girl" and "holding" in Fig. 4. Each row represents a weight vector, and the green and yellow colours denote the highest and lowest score respectively. We observe that the amount of information derived from each module highly depends on the types of different words. More information will come from the relational reasoning module when the question involves the relationship of visual objects. In Fig. 4, different shades of gate values reveals the amount of information derived from each module. When predicting an object word, the relationship weight vector will significantly reduce weight. When predicting the relationship word, it can be seen that the relationship vector is obviously activated. It indicates that the context gate can adaptively control the contribution of the visual feature and the visual relationship feature when predicting different types of words.

V. Conclusion

In this paper, we propose a visual relational reasoning model for image caption. The key of our work is to propose a visual relational reasoning module to explicitly modelling and reasoning the relationship about pairs of relevant visual objects. Moreover, a context gate is introduced to dynamically control the contribution of different modules according to the context which allows predicting different words according to different features. We validated the proposed model through Quantitative analysis and achieved state-of-the-art results on the MSCOCO dataset. Further, we visualized the relational reasoning module and context gating, and performed a qualitative analysis to demonstrate the relationship reasoning and feature selection capabilities of the relational reasoning module model.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grant number 51775513.

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[2] J. Lu, C. Xiong, D. Parikh, and 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. 375–383.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and 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.

[4] J. Gu, J. Cai, G. Wang, and T. Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8957–8964.

[6] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[9] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6580–6588.

[10] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

[11] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[12] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 203–212.

[13] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 684–699.

[14] 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," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5179–5188.

[17] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7190–7198.

[18] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[21] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *Proceedings of the sixth workshop on statistical machine translation*. Association for Computational Linguistics, 2011, pp. 85–91.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[23] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[28] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7219–7228.