Scene Graph Generation Based on Node-Relation Context Module

Xin Lin, Yonggang Li, Chunping Liu(✉), Yi Ji(✉), and Jianyu Yang

Soochow University, Su Zhou, China
{cpliu, jiyi}@suda.edu.cn

Abstract. For better understanding an image, the relationships between objects can provide valuable spatial information and semantic clues besides recognition of all objects. However, current scene graph generation methods don't effectively exploit the latent visual information in relationships. To dig a better relationship hidden in visual content, we design a node-relation context module for scene graph generation. Firstly, GRU hidden states of the nodes and the edges are used to guide the attention of subject and object regions. Then, together with the hidden states, the attended visual features are fed into a fusion function, which can obtain the final relationship context. Experimental results manifest that our method is competitive with the current methods on Visual Genome dataset.

Keywords: Scene graph · Relationship detection · Visual information
Visual attention

1 Introduction

Nowadays, with deeper understanding of images, classifying and locating objects is not enough for some tasks, such as Visual Question Answering problems [1, 2] and Image Caption [3, 4]. It is important to understand the relationships between the object pairs. Through understanding the relationships we can understand not only the spatial structure but also the semantic relationships. As a result, understanding the relationships can help more precise image retrieval [5], object detection [6] and image understanding problems [7].

With the development of neural networks, there are some fast and accurate object detection models [8–10]. They concentrate to recognize a wide variety of objects and regress their bounding boxes. Fast R-CNN [11] is a classic object detection algorithm. Our paper is also based on it. However, only object categories in the image cannot fully represent the complicated real world.

Besides the categories of multiple objects, relationships in the image can provide rich and semantic information. Lu *et al.* propose to use triple structural language to represent relationships between object pairs, such as <object1-predicate-object2> [12]. Relationship is detected using language priors. However, this work concentrates to detect pair-wise relationship rather than understanding the whole image. Xu *et al.* propose a scene graph generation task based on iterative message passing [13]. They train a model to generate scene graph from an image automatically to represent objects

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11302, pp. 134–145, 2018.
https://doi.org/10.1007/978-3-030-04179-3_12

in the image and the complicated topological relationships between them. However, latent visual information is ignored in their message pooling stage. To overcome this problem, we propose a new edge context message pooling method, in which we reenter visual features and we use node GRUs' (Gated Recurrent Unit) [14] hidden state to guide the attention to attend to the more important regions in the corresponding relations. In this way, latent visual information in relationships can be obtained.

On the other hand, the contribution of subject and object in sentence comprehension is not balanced in linguistics [22]. For Example, in the phrase "person holding cup", subject "person" is more important. When do predicate classification, if the subject is "table", "under" or "above" became essential. Based on this assumption, we do importance measure after visual attention when computing edge context message.

In summary, based on Xu *et al.* s' remarkable work [13], we design a node-relation context module. The method consists of two parts: node states guided relation attention module and a better fusion function. In the node states guided relation attention module, we use node GRUs' hidden state guided attention to better utilize the ignored latent visual information. In the fusion function, we measure the contribution of the subject and objects' visual information as well as the hidden states in both edge message pooling and node message pooling to obtain better context messages.

2 Related Work

2.1 Baseline Scene Graph Generation

Xu *et al.* [13] pass context message through node GRUs and edge GRUs iteratively so that the prediction of objects and relationships can benefit from its neighboring context. In the scene graph topology, a relationship triple consists of two node and one edge. For a node, there are inbound edges and outbound edges. For an edge GRU, it receives messages from its neighboring node GRUs. And for a node GRU, it receives message pooling from its inbound and outbound edge GRUs. The module to generate node messages and edge messages is called message pooling. The overview frame work is shown in Fig. 1.

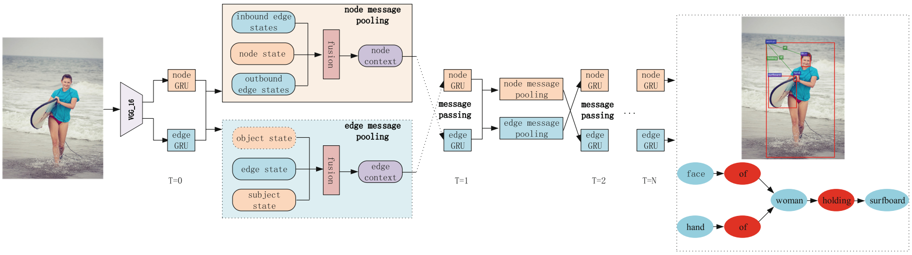

Fig. 1. The architecture of scene graph generation by iterative message passing [13].

Given an image, convolution feature maps are first extracted. Afterwards, the Region Proposal Network [11] generates the region proposals. Using the region proposals, node and relationship feature maps are extracted using ROI-pooling method and fed into corresponding GRUs. Afterwards, the hidden states are fed into message pooling module to generate context message. After a few message passing iterations, a scene graph is generated consisting of object categories, relationships and bounding boxes. The training process is a multi-task learning [15].

2.2 Attention Models

Since Xu *et al.* [13] focused on hidden states formed context, initial visual information is ignored during edge message pooling stage. Instead of treating all feature maps equally, attention mechanism tries to discover different weights of image regions according to their value. Soft attention has been widely used in machine translation [16] and image captioning [3]. In the soft attention mechanism, the hidden state of the LSTM(Long Short-Term Memory) [17] is used to guide attention. In our work, we use the processed node GRU hidden states to guide the attention of the corresponding feature maps to obtain better expression of edge context. To the best of our knowledge, it’s the first time to apply attention module in scene graph generation.

2.3 Relationship Referring

Natural language referring expression is presented and let the model tag the objects referring [18, 19]. Position information, color, object classes and etc. are needed. Recently, structural language is used to referring the object pairs engaged in the relationship expression [12, 20, 21]. Using structural language can reduce the cost of understanding the natural language so that visual understanding task can be focused on. Attention shift algorithm is applied to model relationship in this work. Relationship referring task is a reverse task of scene graph generation. Scene graph generation is a more complicated task since we need to understand all the relationships existing in the image.

3 Node-Relation Context Module

Different from the edge message pooling method in [13], we bring back the ignored visual features and generate a better expression of context. The structure of our module is shown in Fig. 2.

Different from the baseline model, the convolutional feature maps of the corresponding nodes (subject and object) are fed into the edge message pooling module. Moreover, the processed hidden state is used to guide the attention generation. At last, both hidden state context and visual context are fused to generate the final edge context. The node-relation context module consists of two parts: node state guided relation attention module and fusion function. We will present our method from these two aspects.

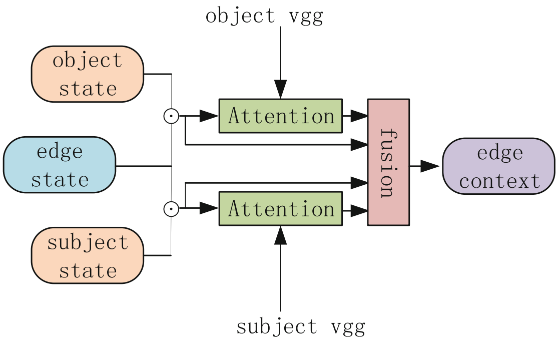

Fig. 2. Illustration of the node-relation context module

3.1 Attention Module for Relationship Prediction

Considering different regions of the object pairs have different contribution to relationship classification, the hidden states of object and subject are used to guide the attention. The concrete framework is shown in Fig. 3 where, $h_{i,t}$ represents the hidden state of $node_i$ at time step t and $h_{i+j,t}$ represents the hidden state of edge GRU relating $node_i$ and $node_j$ at time step t. When $node_i$ and $node_j$ have different characters in the relationship, $h_{i+j,t}$ has different meanings, as expressed in Eq. 1.

$$h_{i+j,t} = \begin{cases} h_{i \to j,t}, sub(i) = True \\ h_{j \to i,t}, else \end{cases} \tag{1}$$

where the decision function $sub(i)$ decides if $node_i$ is subject, if true then true, otherwise false.

We first element-wise multiply the hidden state of node GRU and edge GRU as the guidance of the attention, shown in Eq. 2. Moreover, $s_{i,t}$ is also taken as components of hidden state context to embed the final context expressed in Sect. 3.2.

$$s_{i,t} = h_{i,t} \odot h_{i+j,t} \tag{2}$$

Here, the same as the baseline model, images are fed into a VGG-16 [4] convNet. The convolution feature maps from the last conv layer (i.e., conv5_3) are used. Roi shifts are fed into the model to obtain the feature maps of $node_i$ through roi pooling function, denoted as V_i. V_i is composed of k small regions, $V_i = (v_1, v_{2,} \ldots, v_k)$.

Then, $s_{i,t}$ indicates how much attention the module is placing on different regions. A softmax function is used to gain attention distribution.

$$\alpha_t = \text{softmax}(s_{i,t} \odot V_i) \tag{3}$$

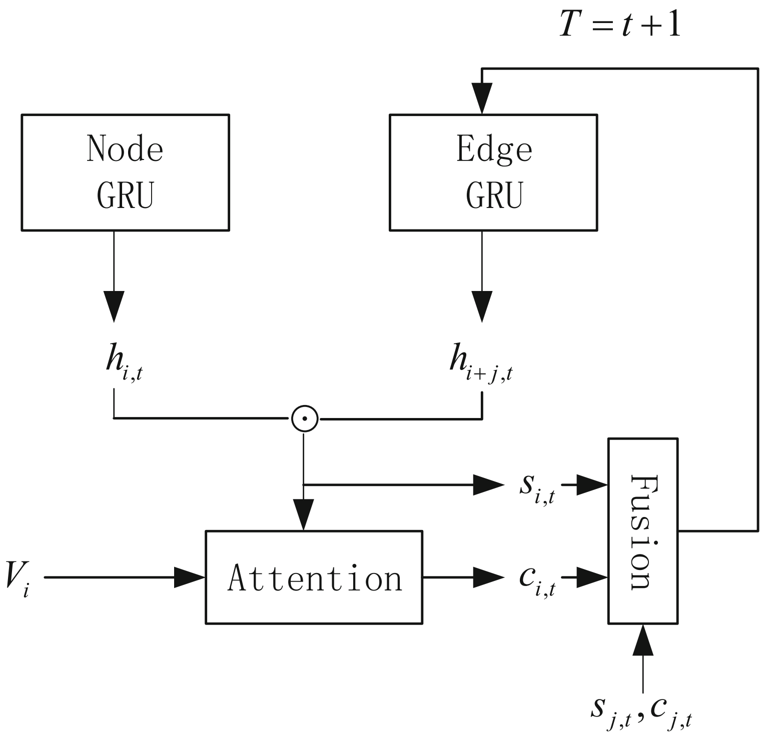

Fig. 3. Node state guided relation attention

Based on the attention distribution, the weighted node feature can be gained by:

$$c_{i,t} = \sum_{l=1}^{k} \alpha_{tl} v_{tl} \tag{4}$$

Due to different context, even the same $node_i$ has different attention distribution in different relationship.

3.2 Fusion Function

After $s_{i,t}$ and $c_{i,t}$ obtained, we will fuse the context vectors. First, contributions of the subject and object in the corresponding relation are computed.

$$sub\_score = \tanh(W_s s_{i,t} + b_s) \tag{5}$$

$$obj\_score = \tanh(W_o s_{j,t} + b_o) \tag{6}$$

Specially, tanh function is also used in node message pooling stage instead. The advantage of this function will be proved in ablation study (Sect. 4.2).

Then, we fuse the subject and object components, shown in Eqs. 7 and 8:

$$v\_c_{i+j,t} = sub\_score \bullet c_{i,t} + obj\_score \bullet c_{j,t} \tag{7}$$

$$g\_c_{i+j,t} = sub\_score \bullet s_{i,t} + obj\_score \bullet s_{j,t} \tag{8}$$

After weight fusion, the hidden state context $g\_c_{i+j,t}$ and the visual context $v\_c_{i+j,t}$ are concatenated and fed into a fully connected layer, as the final context of this relationship.

4 Experiments

Applying our module to the baseline model [13], we can generate better context message and obtain more accurate scene graph. We conduct experiments on large-scale benchmark: Visual Genome [20]. This dataset is a human-annotated scene graph dataset, containing 108,077 images. Each image involves 25 objects and 22 relationships on average. In this section, we analyze our model in four parts: evaluation metrics, ablation study, comparison with existing works and qualitative analysis. The ablation study includes **Baseline** model [13], **Baseline + V**(visual-context), **Baseline + V**(visual-context) **+ H**(hidden-states-context, not sharing the weights), **Baseline + V**(visual-context) **+ H**(hidden-states-context) **+ SW** (sharing weights, our proposed final model, short as ours).

4.1 Evaluation Metrics

Given an image, the scene graph generation task includes locating the objects, predicting their categories and figuring out the relationship between each object pairs. According to the metrics in [12], the evaluation is divided into three levels.

PREDCLS(predicate classification): Given the locations and the categories of the objects, relationships between object pairs are to be predicted.

SGCLs(scene graph classification): Given the bounding boxes of the objects, categories of the objects and the relationships between them are to be predicted.

SGGen(scene graph generation): Locations, categories of the objects and the relationships between the objects all need to be predicted

The difficulty level of these three tasks is from easy to difficult. Image-wise recall evaluation metrics, R@50 and R@100, are adopted to evaluate the three tasks. R@x is abbreviation of Recall@x, meaning the fraction of the ground truth relationships prediction among top x predictions in an image. Obviously, the bigger fraction of the ground truth relationships prediction among top x predictions, the better the model performs. The reight prediction means to classify the triple structure.

4.2 Ablation Study

According the improvement on model message passing [13], we design module analysis experiments to prove the superiority of our node-relation context module. In this subsection, we perform ablative studies to analyze the contribution of each improvement to our module, shown in Table 1.

Table 1. Evaluation results of the contribution of each improvement to our module.

Methods	PREDCLs		SGCls		SGGen	
	R@50	R@100	R@50	R@100	R@50	R@100
Baseline [13]	44.75	53.08	21.72	24.38	3.44	4.24
Baseline + V	43.09	50.96	22.12	24.69	4.36	5.53
Baseline + H	**45.83**	**53.93**	22.25	24.96	**4.51**	**5.61**
Baseline + V + H	45.37	53.61	22.37	25.05	4.18	5.39
Baseline + V + H + SW(ours)	45.55	53.66	**23.37**	**26.29**	4.33	5.57

As we can see, though only visual context model does not perform well in task PREDCLs, it performs well in the other two tasks. Only H-states context model performs best in task PREDCLs and SGGen. Combining these two context, we obtain a model performing best in SGCls. Though a bit of poorer than H-states context model in task PREDCLs and SGGen, considering relatively large exceeding in task SGCls, we make it our final model. As shown in Table 1, sharing the weights between the visual context and hidden states context performs better. In addition, one benefit is achieved that we can save calculation amount obviously. Our model performs best in SGCls task and performs second best in other two tasks. Since there is a larger improvement in SGCls task, we use Baseline + V + H + SW as our final model

To better prove the advantage of tanh function to weigh the contribution of subject and object components, we design a comparison of some common activation functions, including sigmoid function, ReLU (Rectified Linear Unit) and tanh, on both the baseline model and our model. As shown in Table 2, tanh function performs best in both the baseline models and our improved models. Due to the improvements proposed, our model using tanh function performs better.

4.3 Comparison with Existing Works

Table 3 shows the performance of our model against two existing ones. We can see that in each task, our proposed model has exceeded the existing two models. Specially, there is a relatively large increase in task SGCls. Categories of the objects are not offered in task SGCls. Thanks to re-entering the visual features, more valuable information is gained to classify the objects and the relationships.

Table 2. Evaluation results of different fusion functions, including tanh, sigmoid and relu function to calculate weights of subject component and object component.

Methods	Fusion function	PREDCLs		SGCls		SGGen	
		R@50	R@100	R@50	R@100	R@50	R@100
Baseline [13]	Sigmoid	44.75	53.08	21.72	24.38	3.44	4.24
Baseline + r	ReLU	34.03	43.32	18.34	21.37	3.29	4.28
Baseline + t	tanh	45.54	**53.96**	21.89	24.62	4.36	5.43
Ours + s	Sigmoid	44.37	51.97	21.97	24.58	4.36	5.47
Ours + r	ReLU	44.22	52.19	22.21	24.85	**4.38**	5.48
Ours	tanh	**45.55**	53.66	**23.37**	**26.29**	4.33	**5.57**

Table 3. Comparison with two existing models.

Methods	PREDCLs		SGCls		SGGen	
	R@50	R@100	R@50	R@100	R@50	R@100
Language priors [12]	27.88	35.04	11.79	14.11	0.47	0.32
Message passing [13]	44.75	53.08	21.72	24.38	3.44	4.24
Ours	**45.55**	**53.66**	**23.37**	**26.29**	**4.33**	**5.57**

4.4 Qualitative Analysis

In this section, part of the experimental results are shown below.

Figure 4 shows qualitative results using human annotated bounding boxes. The results show that the baseline model confuses about the categories of the objects. For example, it predicts the head of a cute owl <sheep-of-bird>, because the head looks fury. Our model predicts <head-of-bird> instead. Since we obtain better representation of context, it's easier for model to distinguish the categories having similar appearance. What is more interesting, our model predicts the man play tennis wearing "short" and the baseline model predicts "pant". As we know, male tennis players usually wear short and our model can infer this from the context.

Figure 5 shows the results using the bounding box produced by RPN. Our model is par with the baseline model. Using the region proposals, the models produce more reasonable answers than using human annotated even if there are less objects. That's because the RPN network and the generation model share the same feature maps and understand the images from similar angle. Compared with the results generated by human annotated bounding boxes, this model tends to predict uncertain predicates such as "has".

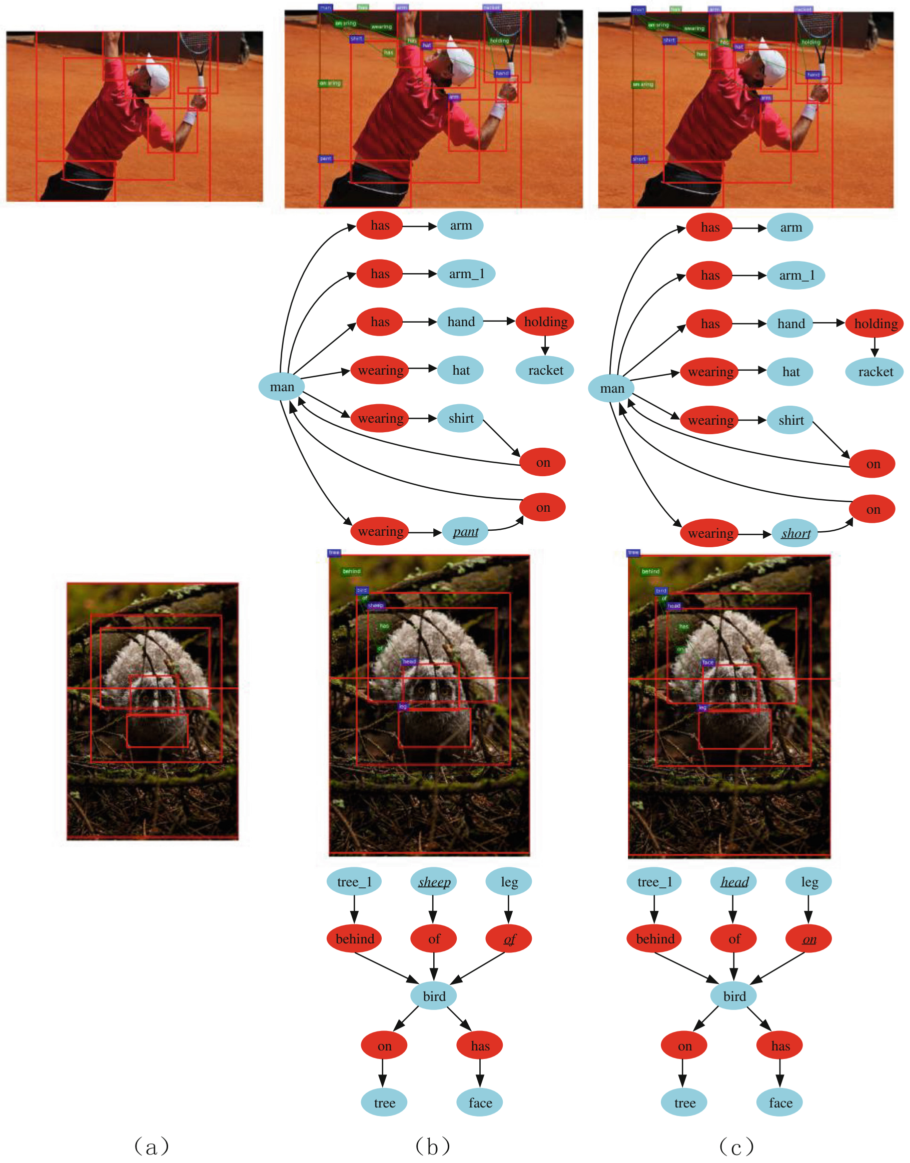

Fig. 4. Part of results from model message passing [13] and our final model, where the (a) is human annotated bounding boxes, (b) are the results of the original model and (c) are the results of our model. The 1st and the 3rd lines are the visualized results over the images and the 2nd and the 4th lines are the final scene graph. The baseline model is retrained by us using the default parameters.

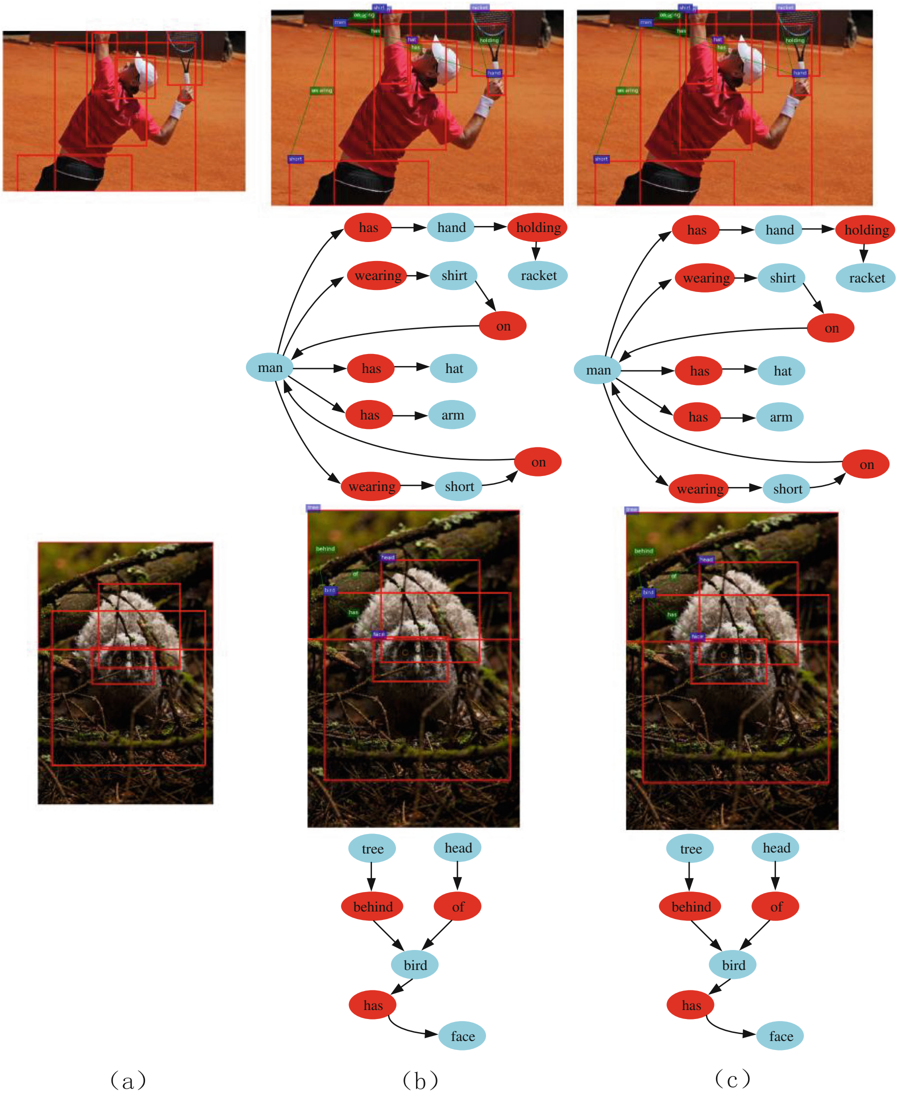

Fig. 5. Part of results from model message passing [13] and our final model, where the (a) is the bounding boxes generated by RPN, (b) are the results of the original model and (c) are the results of our model. The 1st and the 3rd lines are the visualized results over the images and the 2nd and the 4th lines are the final scene graph.

5 Conclusions

In this paper, we propose node-relation context module, improving the performance of scene graph generation. Through introducing the attention guided visual features, we can find latent visual information in relationship. In addition, by using a better fusion function and sharing the weights between subject and object components, we obtain the importance of subject and object context components. However, during the research, we find that the Visual Genome dataset is severely unbalanced in relationship categories and need further research.

Acknowledgement. This work was partially supported by National Natural Science Foundation of China (NSFC Grant No. 61773272, 61272258, 61301299, 61572085, 61272005), Science and Education Innovation based Cloud Data fusion Foundation of Science and Technology Development Center of Education Ministry (2017B03112), Six talent peaks Project in Jiangsu Province (DZXX-027), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), and Provincial Key Laboratory for Computer Information Processing Technology, Soochow University.

References

1. Jang, Y., Song, Y., Yu, Y., et al.: TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1359–1367 (2017)
2. Vinyals, O., Toshev, A., Bengio, S., et al.: Show and tell: a neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
3. Xu, K., Ba, J., Kiros, R., et al.: Show, Attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
4. Simonyan K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
5. Johnson, J., Krishna, R., Stark, M., et al.: Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3668–3678 (2015)
6. Sadeghi, M. A., & Farhadi, A.: Recognition using visual phrases. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1745–1752
7. Li, Y., Ouyang, W., Zhou, B., et al.: Scene graph generation from objects, phrases and region captions. In: IEEE International Conference on Computer Vision, pp. 1270–1279 (2017)
8. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
9. Wang, Z., Chen, T., Li, G., et al.: Multi-label image recognition by recurrently discovering attentional regions. In: IEEE International Conference on Computer Vision, pp. 464–472 (2017)
10. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: a unified framework for multi-label image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)
11. Girshick, R. B.: Fast R-CNN. In: International Conference on Computer Vision, pp. 1440–1448 (2015)

12. Lu, C., Krishna, R., Bernstein, M.S., et al.: Visual relationship detection with language priors. In: European Conference on Computer Vision, pp. 852–869 (2016)
13. Xu, D., Zhu, Y., Choy, C.B., et al.: Scene graph generation by iterative message passing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3097–3106 (2017)
14. Dey, R., Salemt, F.M.: Gate-variants of gated recurrent unit (GRU) neural networks. In: International Midwest Symposium on Circuits and Systems, pp. 1597–1600 (2017)
15. Xue, Y., Liao, X., Carin, L., et al.: Multi-task learning for classification with Dirichlet process priors, **8**(1), 35–63 (2007)
16. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (2015)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
18. Hu, R., Rohrbach, M., Andreas, J., et al.: Modeling relationships in referential expressions with compositional modular networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1115–1124 (2016)
19. Yu, L., Lin, Z., Shen, X., et al.: MAttNet: modular attention network for referring expression comprehension. arXivpreprint: 1801.08186 (2018)
20. Krishna, R., Zhu, Y., Groth, O., et al.: Visual genome: connecting language and vision using crowdsourced dense image an-notations. Int. J. Comput. Vision **123**(1), 32–73 (2016)
21. Krishna, R., Chami, I., Bernstein, M., et al. Referring relationship. arXivpreprint: 1803.10362 (2018)
22. Kibrik, A.E.: Beyond subject and object: toward a comprehensive relational typology. Linguist. Typology **1**(3), 279–346 (1997)