Scene graph generation based on global embedding and contextual fusion

Zhiyong Zhao*, Ronglin Hu, Hongtai Ma, Xinxin Zhang
School of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, Jiangsu, China

ABSTRACT

As a structured abstraction method for objects and their interactions in visual scene, scene graph captures entities in the scene and the relationships between the entity pairs, and helps in understanding the visual scene better. Currently, scene graphs in most research works are generated by modeling using context information among targets, focusing only on the inference process but ignoring the integrity of the input target information and the impact of the global information of the targets on relationship inference. Therefore, a new scene graph generation method based on global embedding and contextual fusion (GECF) is proposed in this paper. In this method, richer entity information is obtained by embedding global information into the entity features, while more robust inference of entity interaction information and more reasonable relationship fusion are acquired by combining the attention weighting module and the context inference module as a joint inference module, and merging the obtained entity features according to their discrepancy. The experiment on Visual Genome dataset shows that GECF method performs better than the existing methods in scene graph visual relationship detection.

Keywords: Scene graph, context inference, global embedding, visual scene understanding

1. INTRODUCTION

Most common computational tasks in scene understanding for computer vision are mainly image classification[1], target recognition[2], and semantic segmentation[3]. With the continuous development of various data-driven based modeling methods in recent years, the accuracy of most of the above tasks has surpassed the manual identification level, except for some complex tasks such as scene semantic understanding[4]. Therefore, the concept of scene graph[5] is proposed for the representation of semantic information by abstracting the images in a structured form based on the instance relationships in scenes.

A scene graph is a high-level graph structured representation of the content in an image, consisting of nodes and connecting edges. Nodes are the entities in the image and edges are the relationships among the entities. As shown in Figure 1, the entities in the scene and the relationships between entities pairs are captured.

At present, a two-stage strategy is adopted in most of the scene graph generation missions: First stage is the entity recognition detection by target detection methods like Faster-RCNN[2], YOLO[6] on scene graph; Second stage is the generation of a graph with triplet of subject-predicate-object by jointly reasoning the generated target information.

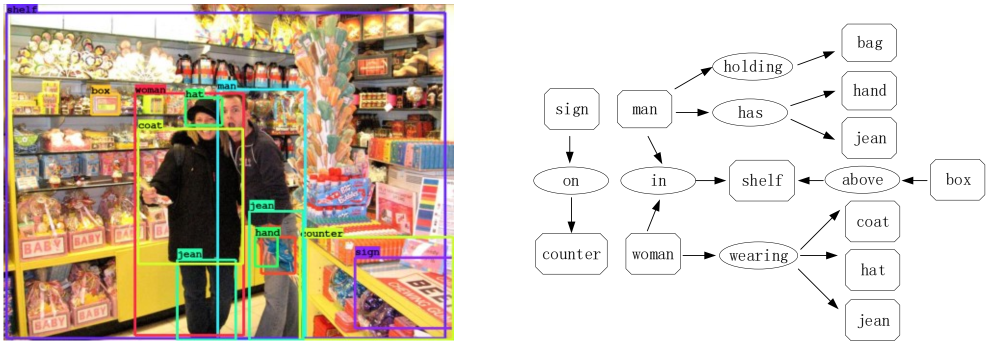
Compared to image understanding tasks such as single target detection, the scene graph contains richer and more abstract semantic information, which can be widely used in visual task applications like image retrieval[5] (retrieving related images by key high-dimensional semantic information of an image), visual question and answer (VQA) based on image information[7, 8], and has also demonstrated its potential in image generation[9] and scene description[10]. With the development of applications in more directions, in order to improve the accuracy of scene graph generation and the robustness of model, some research works focus mainly on two directions: the problem of long-tail distribution in the presence of datasets and modeling based on target information.

Bias effects in the dataset long-tail distribution influence greatly on the diversity of relationship detection. Using common sense knowledge as priori guidance, references[11-13] have improved the detection accuracy of multiple relationships, weakened the bias effects, and made the generation of relationships in a more consistent way with human's basic judgment.

[*] zhiyongzhao11@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125063O
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661798

Although there is an improvement in recall rates, it still contradicts with the idea that visual scene graph should focus more on vision. When common sense knowledge is added to guide the relationship generation, the influence of visual features on the relationship space is weakened, which is still a tricky problem for visually integrated common sense inference.



(a) Target detection graph (b) Graph structure representation generated by the target

Figure 1. Scene graph example.

Besides, the main direction of current research is establishing information transfer model between targets on the aspect of relationship modeling of information[14-18]. Currently, the relationship representation of nodes and edges is mainly obtained by context information interaction[14, 17, 18] and message passing[15, 16]. In the two methods, the nodes and edges are set with the same weights, but the importance differences of the primary and secondary targets for scene understanding are ignored to show flexibly the visual content the image expressing. Due to the separate representation of nodes and edges, and only the information between nodes contextually interconnected, the global structural features of the image can not be fully obtained. Thus it challenges for us to establish a model encoding the global information with model inference.

To solve the modeling problem of common sense weakening visual features and target interaction information, this paper proposes a method incorporating each of the target feature into the raw image visual information based on Neural Motifs[18]. The convolution features of the image are fused with the entity information of the proposed network, which enhances the comprehensive expression ability of the model on the relationships between targets and scenes, so that the model can comprehensively learn the interactive information features of nodes. The long-range dependence among targets is obtained by a non-local weighting operation. Structure decoding the embedding information of categories with attention mechanism reduces the redundant information of each node and raises the focus on the main targets in the scene during model inference. And during the decoding relationship generation process, the global information and weighted information are fused for the relationship features with information differences in subject and object semantics, thus the accuracy of relationships gets improved.

2. SCENE GRAPH GENERATION METHOD

With the continuous iterations of scene graph generation methods, researchers have proposed a scene graph generation model based on Recurrent Neural Networks (RNNs)[12, 15, 16, 19] and a context inference model[13, 14, 17].

The RNNs-based message passing model[15] lists all possible entity pairs and proceeds relationship inference using dynamic planning for iterative messaging, and the flexibility of GRU units eliminates the limitations of RNNs training, but also increases the complexity of the model at the same time. A visual phrase-guided messaging structure with a specific messaging flow is proposed in[19], and the fixed path of message exchange during phrase detection is changed by string-parallel combination. Although the broadcast information is aggregated through the flow mechanism, the target information is incomplete in relationship phrase inference and lacks relative location information. In reference[16], the non-maximum suppression method is adopted to filter overlapping phrase regions and all target pairs are introduced together into a spatially weighted message passing inference model. In this study, the number of nodes is reduced by merging subgraphs but the global information is ignored at the same time.

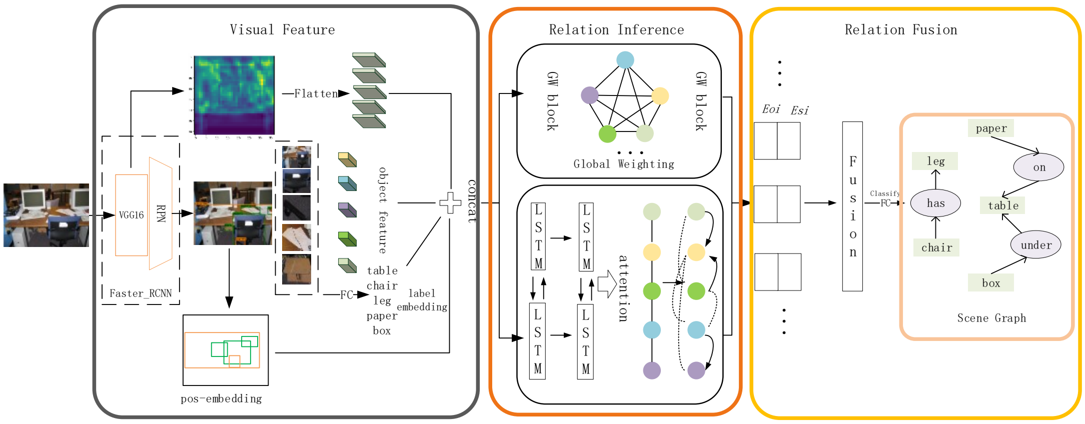


Figure 2. The structure of the scene graph generation model based on global embedding and contextual fusion.

Since the scene as a whole is composed of multiple entities and contexts, the scene graph generation can be represented as a fusion inference of context information. And in most of the inference models, the information is updated mainly on the nodes and edges of the candidate scene graph, and the entity visual features, semantic information and spatial information are taken as the main information for fusion inference. In scene graph generation models, context relationship detection is the major inference. In[14], the obtained entity pairs are filtered in the way from dense to sparse, and then the attention mechanism driven graph convolutional network is applied for context information transfer.

In partial relationship inference, the generation of interrelationships between targets can be guided by introducing additional knowledge. The statistical target co-occurrence information for graph neural networks is used to learn the target context feature inference and relationship inference to build graph structures in reference[12]. Introducing the external knowledge base as common sense[13] solves the problem of data set bias to some extent, but also weakens the influence of the overall visual information on relationship inference. Zhang et al.[17] detect the visual relationship based on complete image information using a fully connected end-to-end relationship inference architecture. In their study, transformation vectors are taken as the relationship representations in the low-dimensional space after mapping of entity features. Although there is an improvement in accuracy, context and global information are both ignored.

Different with most of the existing studies, this paper incorporates global visual features and uses an attention-based mechanism for inference. The global visual feature is introduced when we extract the target information, and the abstract visual information of the scene is taken as the feature of relationship inference and embedded into the information space of each target. The long-range dependence information generated by the attention mechanism is fused with the information of the context model. Then the subject-object relationship space is obtained, and the information difference between the subject and object is used for the relationship inference between the targets.

3. GLOBAL EMBEDDING AND CONTEXTUAL FUSION (GECF) MODEL

The scene graph generation framework in this paper contains three sub-modules: visual information generation module, relationship fusion module, and relationship inference module as shown in Figure 2. And the proposed model aims to generate graph structure representations of images, get the entity information during the target detection stage, and incorporate global image high-level features for richer entity features, on which a global weighting module combined with an attention mechanism in a long-range dependency module is used. Then the global context interaction information is captured and the differences of the fused information is pointed to the relationship inference.

3.1 Problem definition

In scene graph generation tasks, images are transformed to graph structures by $I \to G$. G, the scene graph structure, is used to describe the target nodes in the scene graph and the predicate relationships between the target nodes, which contain the candidate target box $B=\{b_1, b_2, \ldots, b_n\}$, $b_i \in \mathbb{R}^4$. The category collection corresponding to the target box is $O=\{o_1, o_2, \ldots, o_n\}$, $o_i \in C^{151}$, where C^{151} is the candidate category collection truth set. The triadic relationship set between nodes is $R=\{r_1, r_2, \ldots, r_m\}$, where r_k is the subject-predicate-object triplet, with subject node (b_i, o_i), $(b_j, o_j) \in B \times O$; predicate node $p_{i \to j} \in P^{51}$, P^{51} as the candidate relationship truth set. Therefore, the scene graph generation process containing image I to graph structure G, and target box B, target category O and subject-object relationship R can be expressed by the following factorization model:

$$P_r(G|I)=P_r(B|I)P_r(O|B,I)P_r(R|O,B,I) \tag{1}$$

where $P_r(B|I)$ is the probability of the entity candidate box obtained by the target detection model (section 3.2); $P_r(O|B,I)$ is the probability distribution of the entity box categories in target detection; $P_r(R|B,O,I)$ is the probabilities of the inference fusion relationship among image I, category O and candidate target box B (sections 3.3 and 3.4).

3.2 Visual information generation module

For a given image, we use the target detection method for the targets and the convolutional neural networks for the image's advanced feature F^1, $F^1\in\mathbb{R}^{n\times512\times w\times h}$. The alignable fused visual feature $F^{\mathrm{O}}\in\mathbb{R}^{n\times512}$ can be obtained by mean-pooling mapping on F^1, and can be calculated by

$$F^{\mathrm{O}}=AVG(F^1) \tag{2}$$

where AVG takes the mean sampling of the feature F^1 with size of $w\times h$ and transforms them to 1×1. To match with the size of the feature fusion, the matrix $n\times512\times1\times1$ is dimensionally compressed and finally F^{O} is obtained by flattening operation on features mapping.

The proposed target visual features $V=\{v_1, v_2, \ldots, v_\mathrm{n}\}$, $v_\mathrm{i}\in\mathbb{R}^{4096}$ are obtained through the target detection module, where each target i corresponds to a bounding box b_i, and the candidate bounding box collection $B=\{b_1, b_2, \ldots, b_\mathrm{n}\}$, $b_\mathrm{i}=(x_{i1}, y_{i1}, y_{i2}, y_{i2})$. Since the obtained entity information is regional, it cannot reflect the size relationship between targets.

We try to recompute the bounding box features and construct the following encoding way:

$$pos_{b_i}=fc\left(\left[b_i:\frac{(x_{i1}+x_{i2})}{2}:\frac{(y_{i1}+y_{i2})}{2}:|y_{i1}-y_{i2}|\times|x_{i1}-x_{i2}|\right]\right) \tag{3}$$

where ":" and fc are used as splicing and full connection operations. In order to align the input fusion features and the extensive information space, equation (3) has full connection adjustment on the code information and then $pos_{b_i}\in\mathbb{R}^{n\times128}$ is obtained, and so as the spatial relative size of each target.

Lastly, we take all the above information into consideration and get the overall visual information F_i^{in} of the ith node as follows:

$$F_i^{in}=\left[v_i:S_{oi}:pos_{b_i}:F_i^{\mathrm{O}}\right] \tag{4}$$

where F_i^{O} is the visual information of the ith node, and the category embedding feature S_{oi} is obtained by embedding words into GloVe[20].

3.3 Relationship inference module

The relationship inference module in GECF model contains two sub-modules: attention structure based context module and long and short-term memory networks based context module. The two sub-modules map the target information in different ways using the before-and-after information of sequence data to the relationship space in parallel, so that the information transferring among nodes and edges can be sufficiently proceeded and the long-range dependence between the targets can be captured.

3.3.1 Attention Structure Based Context Module. In the area of context information modeling, Cao et al.[21] propose a simplified GC (Global Context) module as shown in Figure 3a. The excessive operation volume caused by nonlocal weighting structures drops using the GC module[22], which also solves the problem of the insufficient effectiveness of the channel attention mechanism on global context modeling[23].

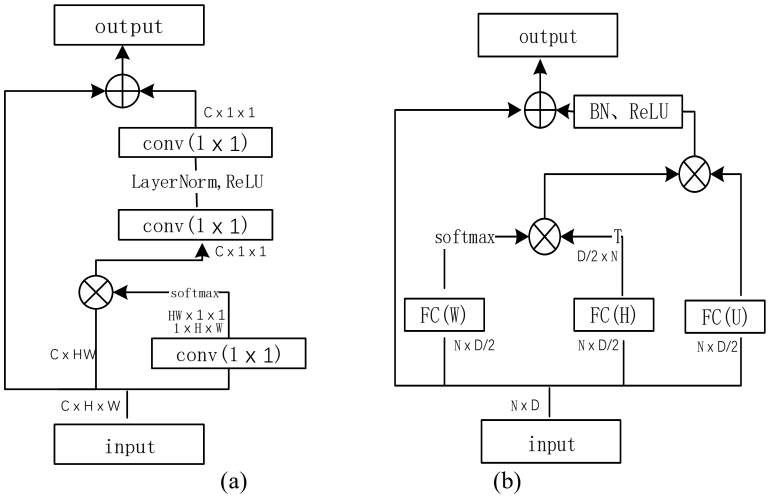


Figure 3. Attention-based global context module.

Inspired by the Non-local[22] and SENet[23] methods, this paper also introduces a global context method into relationship inference and constructs a global weighting module as shown in Figure 3b, in which the long-range dependency relationship between targets is captured after using the multiple target features in an image to iteratively update the adjacency information of a single target feature.

The fused information in equation (4) is encoded using the global weighting module in Figure 3b, among which the single weighting module is calculated as follows:

$$C_1 = \sigma(WF^{in})(HF^{in})^T \tag{5}$$

$$C_2 = F^{in} + fc(C_1(UF^{in})) \tag{6}$$

where σ is the SoftMax function; *W*, *H*, *U* are learnable parameters for the fully connected layer; *fc* is the fully connected operation.

3.3.2. Long Short-Term Memory Networks Based Context Module. Based on the bidirectional LSTM network, context information transfer in MOTIFS[18] is carried out. But the study ignores the important information in the key nodes and edges, and the generated scene graph is with ambiguous main relationships and unclear key targets. BiLSTM-Attention network is applied to improve the relationship inference in this paper inspired by the relation extraction in natural language processing[24, 25]. Attention mechanism based model focuses more on the primary relationship between the targets and obtains the information representation of key nodes by following the hidden layer information of the neighboring target nodes. The hidden layer feature information is as follows:

$$A_1 = \text{BiLSTM}(F^{in}) \tag{7}$$

where A_1 is the hidden layer feature information of the target node after encoding. The target information A_2 in the context of each node feature is

$$A_2 = A_1\sigma(\tanh(KA_1)Q) \tag{8}$$

where *Q* and *K* are randomly initialized trainable parameters, and tanh is the activation function. To weigh the importance of a node, *Q* is used as a similarity measure represented by advanced query representation. The context relationship information space is obtained by multiplying A_1 with the normalized weight matrix.

3.4 Relationship fusion module

We decode the relational spatial information of the inference module by the bidirectional LSTM and get the context features of the edges:

$$[E_o{:}E_s]=\mathrm{BiLSTM}(fc([C_2{:}A_2])) \tag{9}$$

where E_o and E_s are the semantic features of the subject and object, $E_o, E_s \in \mathbb{R}^{4096}$. Subject and object features are included in the obtained context space. In previous studies, the probability of relationship classification is mainly obtained by multiplying the separated subject and object features. We can see from reference[26] that nonlinear projection on fusing visual feature *x* and problematic feature *y* to measure the difference between them. The equation is followed:

$$x\Diamond y=\mathrm{ReLU}(W_x x+W_y y)-(W_x x-W_y y)^2 \tag{10}$$

where ReLU is the activation function; W_x, W_y are learnable parameters. Similarly in this paper, we take different information of the subject and object to measure the probability distribution of their corresponding predicate relations, which is called R_i and calculated as follows:

$$R_i=\max(E_o+E_s)-(E_o-E_s)^2 \tag{11}$$

4. EXPERIMENTS AND RESULTS

The effectiveness of the GECF model is verified based on Visual Genome dataset. And contrast and ablation analysis for the modules in the model are proceeded. The inference results of the model are evaluated based on three subtasks of predicate classification PredCls, scene graph classification SGCls and scene graph generation SGGen, respectively.

4.1 Dataset and evaluation indicators

In this paper, we use the same dataset Visual Genome and evaluation indicators as references[14, 18, 27, 28]. Specifically, considering the influence of long-tail distribution, 150 common object categories and 50 relationship categories are selected for evaluation. After the dataset pre-processing, the scene graph for each image has an average of 11.6 objects and 6.2 relationships. For the comparison with existing methods, the dataset is separated into a training set, a validation set, and a test set. The training set contains 75651 images, in which 5000 images are in the validation set, and rest of the 32422 images are in the test set. Three subtasks are set to evaluate the effect of the scene graph generation model:

PredCls (Predicate Classification): set the bounding borders and labels of the correct location of entities; classify relationships between the targets;

SGCls (Scene Graph Classification): set the bounding borders of the targets; predict the labels of the targets in the border, and then classify the relationships between the target pairs;

SGGen (Scene Graph Generation): take only one original image; detect the targets in the image for their bounding border information and labels; classify the relationship between the target pairs.

In this paper, we use the recall rate Top-K as evaluation indicator, denoted as Recall@K, which represents the proportion of correctly predicted classifications in the first K predicted relationships and K is set to 20, 50 and 100 respectively.

4.2 Experimental setting

A two-stage training method is applied in the scene graph generation in this paper. We use the Faster-RCNN target detection model pre-trained by VGG16[29] for entity detection. Images are all with a uniform size of 592 × 592, and normalized with mean and variance. When we get the global information, the abstract visual feature map is obtained by mean-pooling and full connection on VGG16 output. The GloVe[20] model is used to convert the categories into word vectors at the time of category embedding, and a fully connected layer with an output dimension of 256 is applied for the fusion of the output features of the self-attentive structure and the bidirectional LSTM. The stochastic gradient descent (SGD) is set as the convergence algorithm, with the parameter batch set to 1, the base learning rate to 0.001, the weight decay to 0.0001, and the momentum to 0.9. The model is trained on RTX-2080tiGPU based on PyTorch framework for single card training.

4.3 Quantitative analysis

As shown in table 1, the method proposed in this paper containing three subtasks on Visual Genome dataset are compared with the scene graph generation methods of IMP[15], AE[30], TFR[31], G-RCNN[14], and MOTIFS[18]. We use the same target detection pre-trained model for each method. The experiment results show that the GFCF model in this paper performs better in handling the three subtasks than the other methods. As to PredCls, the recall rates of Recall@20\50\100 are 61.6%

，66.9% and 68.4% respectively, with an improvement of more than 1%. This certifies that the model has a good performance on relationship prediction classification, especially in SGCLs.

Table 1. Recall rates of GECF and the existing models.

Model	**PredCls Recall@**			**SGCls Recall@**			**SGGen Recall@**		
	20	**50**	**100**	**20**	**50**	**100**	**20**	**50**	**100**
IMP+[15, 18]	52.7	59.3	61.3	31.7	34.6	35.4	14.6	20.7	24.5
AE[30]	47.9	54，1	55.4	18.2	21.8	22.6	6.5	8.1	8.2
TFR[31]	40.1	51.9	58.3	19.6	24.3	26.6	3.4	4.8	6.0
GRCNN[14]		54.2	59.1		29.6	31.6		11.4	13.7
MOTIFS[18]	58.5	65.2	67.1	32.9	35.8	36.5	21.4	27.2	30.3
GECF	61.6	66.9	68.4	37.9	40.6	41.3	22.3	27.7	30.5

4.4 Ablation experiments

In order to obtain the specific impact of the method proposed in this paper on scene graph generation, we design ablation experiments on embedding global information, introducing attention mechanisms and the fusion way in the final relationship inference respectively.

4.4.1 Global Embedding Module Analysis. In order to verify the overall impact of global information on scene graph generation, the attention mechanism based network structure is adopted as the model benchmark, and the global information is embedded with multiples of 1, 1.5 and 2, and the importance is weighted before information fusion. The influence of global information with different weights on relationship classification are shown in table 2. We compare the performances of four weights on PredCls and certify that the weight of global information has a certain impact on the final result. With no global information embedding (V=0) as the benchmark, we enlarge the weight gradually. When V=2, the main influence of the target on relationship inference is reduced due to the model over-fitting visual information. The experiments show that the best performance is achieved when V=1.5, and a moderate amount of information embedding is helpful for the relationship classification and able to integrate environmental information into inference at the time the target information dissemination. The overall results show that the embedding of global information contributes to improving the correct rate of relationship prediction.

Table 2. Recall rates of the global information with different weight values.

V	**PredCls Recall@**		
	20	**50**	**100**
0	59.7	64.7	66.3
1	60.4	65.8	67.4
1.5	61.6	66.9	68.4
2	59.4	64.4	65.9

4.4.2 Relationship Fusion Inference Module. Table 3 shows the comparison result of the attention mechanism structure (BiLSTM+A) based method in this paper and the benchmark model BiLSTM[18]. Since the target category and bounding box are determined at the time of relationship classification, during information transmission, the attention mechanism structure does not act well in extracting the relationship features, especially the visual and category-embedded features but does improve slightly in relationship classification accuracy. But for the scene classification task, several evaluation indicators are improved by more than 3%. Only entity bounding boxes are generated in the target detection stage, and the target classification probabilities are obtained by combining neighboring target features when the model performs information inference, making the target classification more fault-tolerant and extensive. Moreover, attention mechanism structure makes it possible to obtain key semantic information when discriminating target classification and then facilitate scene classification.

Table 3. Recall rates of BiLSTM and BiLSTM+A.

Method	PredCls Recall@			SGCls Recall@		
	20	50	100	20	50	100
BiLSTM[18]	58.5	65.2	67.1	32.9	35.8	36.5
BiLSTM+A	60.6	65.1	67.6	35.9	38.5	39.2

We proceed different combinations of context inference (CI) and global weight (GW) modules for the verification of feature concatenate (Cat) and feature summary (Sum). Mul and ReLU are the two computing methods of element-by-element multiplication of the subject and object, and nonlinear activation during relationship inference. In this paper, we verify the performance of subtask PredCls and the results are shown in Table 4.

In Table 4, compared with the benchmark context inference model, the embedded global weighting module helps in improving the accuracy greatly by enriching semantic features. The influence of different connection methods of feature information on the accuracy is acquired by the comparison of two fusion methods: concatenate and summary. The experimental results show that the summary makes the relationship classification accuracy decrease, and the accuracy is higher with the concatenate. It verifies that with the increase of features contributes to the relationship classification. Besides, for the way of subject-object fusion during relationship classification inference, the experimental results show that the accuracy is higher using nonlinear activation.

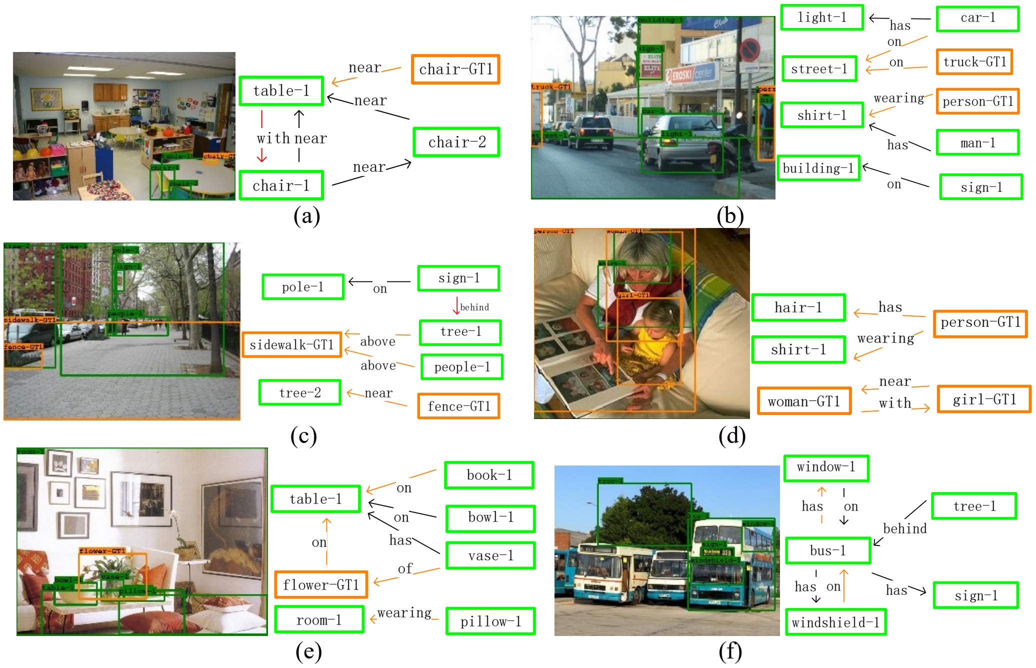
Table 4. Recall rates of different modules.

Exp	Module		FuMethod		Connect		PredClsRecall@	
	CI	GW	Mul	ReLU	Cat	Su	50	100
1	✓		✓				65.1	67.6
2	✓	✓	✓		✓		66.6	68.0
3	✓	✓	✓			✓	65.6	67.6
GECF	✓	✓		✓	✓		66.9	68.4

4.5 Qualitative results

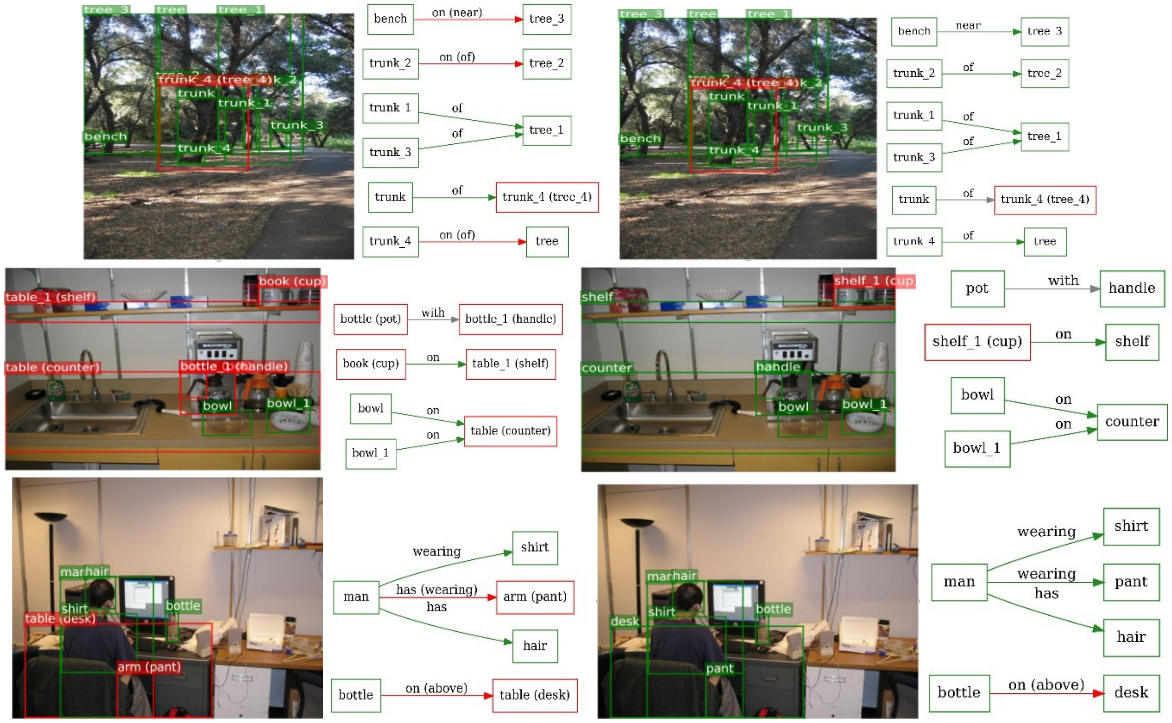
To qualitatively validate the constructed scene graph and visual relationship model, we put some examples of visualizations in the VG dataset based on the SGCls subtask in Figure 4. In the scene images in Figure 4, the green bounding box is the correct prediction zone and matches with the correct label, and the orange bounding box is the correct bounding box without a match. In the scene graphs, the black edge represents the correct predicate classification, the orange includes not only the negation of detected correct predicate, but also negation of relationships between targets that exist but are not detected, while the red represents the wrong predicate prediction classification. For example, in Figure 4a, several different types of targets are correctly detected and relationships are correctly predicted, where the unlabeled <bus-has-window>, <windshield-on-bus> conforming to common sense relationships are also predicted. For targets that cannot be accurately predicted, their associated relationships cannot be predicted either, such as <person-wearing-shirt>, <girl-near-woman> in Figure 4e. Besides, we find that there are conflicting predictions on some targets with accurate predictions, like <table-with-chair> and its correct relationship <chair-near-table> shown in Figure 4b and also wrong prediction on non-significant relationship between targets. There are false relationship detections from the model proposed in this paper, most of which are caused by target detector failure and incorrect detection, like <sign-behind-tree> in Figure 4d. On the premise of accurately detecting the targets, the relationship prediction model in this paper still performs good.

To validate the improvement of the model in this paper on qualitative results, the scene graph generation is visualized and compared with part of the qualitative results of Neural[18] method in the SGCls subtask. As shown in Figure 5, green boxes are the correctly predicted target classifications while the red boxes are the incorrectly predicted target classifications, with the correct classification in brackets. Green edges are the correct relationship predictions of the while the red edges are the wrong relationship predictions, with the correct relationship in brackets. In Figure 5, although there are errors in single target and relationships, the overall target detection results and relationship Recognition performance are better than those of the neural method.



(a) (b) (c) (d) (e) (f)

Figure 4. Scene graph visualization.



Neural method GECF method

Figure 5. Comparisons of graph structure visualization between GECF and Neural methods.

5. CONCLUSIONS

We propose a new scene graph generation method based on global embedding and contextual fusion, incorporating the high-level features of images into entity information, capturing global context interaction information between nodes by a global weighting module and a long-range dependency model combining an attention mechanism, and lastly inferencing the relationship representations by the difference in fused information. An extensive comparison and ablation experiments are conducted with the Visual Genome dataset. And the results show that our method performs better than the existing methods for scene graph generation and the effectiveness of our model is also certified.

REFERENCES

[1] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," NIPS, 25, (2012).

[2] Ren, S. Q., He, K. M., Girshick, R. M. and Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149(2016).

[3] Chen, L. C., Zhu, Y. K., Papandreou, G., Schroff, F. and Adam, H., "Encoder-decoder with atrous separable convolution for semantic image segmentation," Proceedings of the European Conference on Computer Vision (ECCV), 801-818(2018).

[4] Ramanathan, V., Li, C. C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C. and Li, F., "Learning semantic relationships for better action retrieval in images," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1100-1109(2015).

[5] Johnson, J., Ranjay, K., Michael, S., Li, L. J., David, S., Michael, B. and Li, F., "Image retrieval using scene graphs," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3668-3678(2015).

[6] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., "You only look once: Unified, real-time object detection," Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 779-788(2016).

[7] Tang, K. H., Zhang, H. W., Wu, B. Y., Luo, W. and Liu, W., "Learning to compose dynamic tree structures for visual contexts," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6619-6628(2019).

[8] Teney, D., Liu, L. and Hengel, A. V. D., "Graph-structured representations for visual question answering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1-9(2017).

[9] Johnson, J., Gupta, A. and Li, F. F., "Image generation from scene graphs," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1219-1228(2018).

[10] Yang, X., Tang, K. H., Zhang, H. W. and Cai, J. F., "Auto-encoding scene graphs for image captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10685-10694(2019).

[11] Lu, C. W., Krishna, R., Bernstein, M. and Li, F., F., "Visual relationship detection with language priors," Proceedings of the European Conference on Computer Vision (ECCV), 852-869(2016).

[12] Chen, T. S., Yu, W. H., Chen, R. Q. and Lin, L., "Knowledge-embedded routing network for scene graph generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6163-6171(2019).

[13] Gu, J., Zhao, H. D., Lin, Z., Li, S., Cai, J. F. and Ling, M. Y., "Scene graph generation with external knowledge and image reconstruction," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1969-1978(2019).

[14] Yang, J. W., Lu, J. S., Lee, S., Batra, D. and Parikh, D., "Graph R-CNN for scene graph generation," Proceedings of the European Conference on Computer Vision (ECCV), 670-685(2018).

[15] Xu, D. F., Zhu, Y. K., Choy, C. B. and Li, F. F., "Scene graph generation by iterative message passing," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5410-5419(2017).

[16] Li, Y. K., Ouyang, W. L., Zhou, B. L., Shi, J. P., Zhang, C. and Wang, X. G., "Factorizable net: An efficient subgraph-based framework for scene graph generation," Proceedings of the European Conference on Computer Vision (ECCV), 335-351(2018).

[17] Zhang, H. W., Kyaw, Z., Chang, S. F. and Chua, T. S., "Visual translation embedding network for visual relation detection," IEEE conference on Computer Vision and Pattern Recognition (CVPR), 5532-5540(2017).

[18] Zellers, R., Yatskar, M., Thomson, S. and Choi, Y. J., "Neural Motifs: Scene graph parsing with global context," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5831-5840(2018).

[19] Li, Y. K., Ouyang, W. L., Wang, X. G. and Tang, X. O., "VIP-CNN: Visual phrase guided convolutional neural network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1347-1356(2017).

[20] Pennington, J., Socher, R. and Manning, C. D., "Glove: Global vectors for word representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543(2014).

[21] Cao, Y., Xu, J. R., Lin, S., Wei, F. Y. and Hu, H., "GCNet: Non-local networks meet squeeze-excitation networks and beyond," Proceedings of the IEEE/CVF International conference on computer vision workshop (ICCVW) (IEEE), 0-0(2019).

[22] Wang, X. L., Girshick, R., Gupta, A. and He, K. M., "Non-local neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2018).

[23] Hu, J., Shen, L. and Sun, G., "Squeeze-and-excitation networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7132-7141(2018).

[24] Zhou, P., Shi, W., Tian, J., Qi, Z. Y., Li, B. C., Hao, H. W. and Xu, B., "Attention-based bidirectional long short-term memory networks for relation classification," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 207-212(2016).

[25] Yang, Z. Z., Yang, D., Dyer, C., He, X. D., Smola, A. and Hovy, E., "Hierarchical attention networks for document classification," Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1480-1489(2016).

[26] Zhang, Y., Hare, J. and Prügel-Bennett, A., "Learning to count objects in natural images for visual question answering," arXiv Preprint arXiv:1802.05766, (2018).

[27] Li, Y. K., Ouyang, W. L., Zhou, B. L., Wang, K. and Wang, X. G., "Scene graph generation from objects, phrases and region captions," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 1261-1270(2017).

[28] Chen, L., Zhang, H. W., Xiao, J., He, X. N., Pu, S. L. and Chang, S. F., "Counterfactual critic multi-agent training for scene graph generation," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4613-4623(2019).

[29] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," Computer Science, (2014).

[30] Newell, A. and Deng, J., "Pixels to graphs by associative embedding," NIPS 30, (2017).

[31] Hwang, S. J., Ravi, S. N., Tao, Z. R., Kim, H. J., Collins, M. D. and Singh, V., "Tensorize, factorize and regularize: Robust visual relationship learning," Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition (CVPR), 1014-1023(2018).