Knowledge-Enhanced Scene Graph Generation with Multimodal Relation Alignment (Student Abstract)

Ze Fu[1,2], Junhao Feng[1,2], Changmeng Zheng[3], Yi Cai[1,2]*

[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]Key Laboratory of Big Data and Intelligent Robot (SCUT), Ministry of Education, China
[3]Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
{seorangefu, msjh_feng}@mail.scut.edu.cn, csczheng@comp.polyu.edu.hk, ycai@scut.edu.cn

Abstract

Existing scene graph generation methods suffer the limitations when the image lacks of sufficient visual contexts. To address this limitation, we propose a knowledge-enhanced scene graph generation model with multimodal relation alignment, which supplements the missing visual contexts by well-aligned textual knowledge. First, we represent the textual information into contextualized knowledge which is guided by the visual objects to enhance the contexts. Furthermore, we align the multimodal relation triplets by co-attention module for better semantics fusion. The experimental results show the effectiveness of our method.

Introduction

Scene Graph Generation (SGG) is a task to provide a graph representation for fine-grained understanding of an image and involves the detection of all *(subject, predicate, object)* triplets in an image. Recent progress (Tang et al. 2020) in scene graph generation addresses the problems of visual reasoning or unbalanced data distribution which only relying on the given image. However, few methods reveal the corresponding relations between visual contexts and textual contents. As a result, most existing models suffer the limitations of lacking of contextual semantics when the surrounding environment is insufficient to indicate the object behaviors. For example, as shown in Figure 1 (right), it is hard to detect the correct predicate *holding* between man and trophy caused by the lost contexts of holding action. Fortunately, we can still know that the man is holding a trophy indicated by the texts in Figure 1 (left), that is, *Kobe Bryant wins his NBA MVP*.

Although textual contents can be utilized to supplement the missing semantics of visual contexts, simple introduction of textual semantics with representation concatenation will bring semantic disparity and inevitably result in a poor performance. Therefore, the semantics alignment between multimodal information should be taken to improve this situation. Some knowledge triplets *(entity, relation, entity)* extracted from texts might be corresponding to some visual objects and their relations, which brings a cross-modality relation mapping. It motivates us to leverage textual knowledge to help visual scene graph generation. As depicted in Figure 1, the textual triplet *(Kobe Bryant, awarded, NBA MVP)* extracted by information extraction model may reflect the visual triplets *(man, holding, trophy)*.

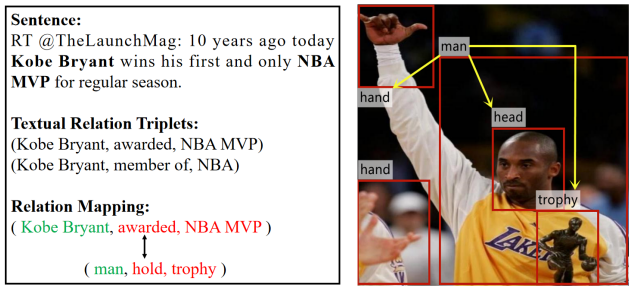

Figure 1: An example of Multimodal Relation Alignment

In this paper, we propose a **k**nowledge-**e**nhanced scene graph generation model with **m**ultimodal **r**elation **a**lignment (KeMRA). We use the knowledge triplets extracted by a pre-trained information extraction model to supplement the missing semantics of visual contexts. To align the visual and textual relations, we apply a co-attention module to capture the correlations. The experimental results show that our model can effectively improve SGG performance with the guidance of textual knowledge and our alignment strategy.

Our Model

Feature Representation Taking an image I and a corresponding text T as input, we first extract the visual object representation by pre-trained Faster R-CNN. The image is transformed into a set of region proposals B with object-level visual features Y, vectors L of object label probabilities. Then we adopt a pre-trained BERT to extract the textual representation X with relation triplets semantics.

Cross-modal Object-level Alignment Module To obtain the textual features which has the relevant semantics corresponding to the objects, we adopt cross-modal attention mechanism to align the semantics of the multi-modal features and achieve the object-level features alignment. The cross-modal attention module can be denoted as below:

$$Y^* = CMA^*(X, Y)$$

After the alignment, the textual-guided visual representation Y^* is generated. Finally, we integrate the information

*Corresponding author: Yi Cai, ycai@scut.edu.cn

of Y^* by summation and concatenate it with each entity pair vector in X to obtain the textual knowledge representation X^* which contains aligned multimodal semantics.

Visual Context Extraction Module We adopt the Stacked Motif Network (Zellers et al. 2018) as the basic model to generate contextualized visual representations for prediction. These visual representation can be denoted as P:

$$P = MotifsNet(B, Y, L)$$

Relation-level Alignment and Prediction To fully exploit the relation triplets information, we also adopt cross-modal attention mechanism to achieve the relation-level semantics alignment. Therefore, we exploit the implicit entity pair semantics to help complete the missing visual contexts in the image. The entity pair features are guided by the object pair features and transform into a matrix E as follows:

$$E = CMA(X^*, P)$$

We fuse the multimodal aligned features by vectors concatenation. Then, we input the final representation to a MLP to classify the predicate category between the object pairs.

$$output = softmax(MLP([P; E]))$$

Experiment

Dataset

MNRE-SG. MNRE (Zheng et al. 2021) is a multimodal relation extraction dataset. We manually annotate the scene graph of images to extend it to MNRE-SG containing scene graph groundtruth with 5724 images and 9741 sentences.

Baseline Methods

We compare our model with several scene graph generation models, including IMP (Xu et al. 2017), Motifs (Zellers et al. 2018), VCTree (Tang et al. 2019).

Experimental Results

We evaluate the SGG performance in two subtasks: (1) Predcls: taking the ground truth object bounding boxes and labels as input, (2) SGcls: taking the ground truth bounding boxes without labels as input, which requires more visual context to detect the scene graph. We use mean Recall @20 proposed by (Tang et al. 2019) to avoid the bias produced by the conventional metrics R@20.

As shown in Table 1, as KeMRA supplements the contexts with sufficient textual knowledge and well aligns the multimodal relation triplets, it greatly improves the performance of our single modality backbone Motifs. Besides, our model achieves the best result compared to other models in two SGG subtasks.

	Predcls	SGcls
Model	mR@20	
IMP	15.53	7.90
Motifs	54.60	30.32
VCTree	59.11	33.11
KeMRA (Ours)	**60.67**	**40.06**

Table 1: Models performance on MNRE-SG *test* split

	Predcls	SGcls
Model	mR@20	
Motifs	54.60	30.32
Motifs + CMA	55.49	35.73
Motifs + BERT	58.90	31.63
Motifs+CMA+BERT (Ours)	**60.67**	**40.06**

Table 2: Results of ablation study

We also conduct the ablation study on our model as shown in Table 2. The variant models without BERT indicate that the textual data is represented by GloVe+LSTM. CMA represents the cross-modal attention alignment mechanism. With the contextualized and well aligned representation obtained by CMA and BERT encoder, our complete model gains significant improvement.

Conclusion

In this paper, we propose KeMRA which supplements the missing visual contexts by well-aligned textual knowledge. Our model builds the cross-modal alignment between multimodal relation triplets. The experimental results depict that the proposed method achieves significant improvement compared to existing models.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62076100), and Fundamental Research Funds for the Central Universities, SCUT (D2210010,D2200150,and D2201300), the Science and Technology Planning Project of Guangdong Province (2020B0101100002)

References

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tang, K.; Zhang, H.; Wu, B.; Luo, W.; and Liu, W. 2019. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural Motifs: Scene Graph Parsing With Global Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, C.; Wu, Z.; Feng, J.; Fu, Z.; and Cai, Y. 2021. MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.