IMPROVE IMAGE CAPTIONING VIA RELATION MODELING

*Feicheng Huang, Zhixin Li**

Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China

ABSTRACT

The performance of image captioning has been significantly improved recently through deep neural network architectures combining with attention mechanisms and reinforcement learning optimization. Exploring visual relationships and interactions between different objects appearing in the image, however, is far from being investigated. In this paper, we present a novel approach that combines scene graphs with Transformer, which we call SGT, to explicitly encode available visual relationships between detected objects. Specifically, we pretrain an scene graph generation model to predict graph representations for images. After that, for each graph node, a Graph Convolutional Network (GCN) is employed to acquire relationship knowledge by aggregating the information of its local neighbors. As we train the captioning model, we feed the potential relation-aware information into the Transformer to generate descriptive sentence. Experiments on the MS COCO dataset validate the superiority of our SGT model, which can realize state-of-the-art results in terms of all the standard evaluation metrics.

***Index Terms*—** image captioning, Transformer, scene graphs, reinforcement learning, attention mechanisms

1. INTRODUCTION

To date, state-of-the-art image captioning methods tend to adopt the encoder-decoder framework derived from neural machine translation to generate corresponding description for a given image. In such a framework, the whole captioning model can be divided into two components, i.e., image understanding and text generation [1] [2]. As a whole, this process is consistent with the end-to-end sequence generation manner.

Usually, attention mechanisms [3] [4] [5] [6] and reinforcement learning [7] [8] [5] are combined with the common encoder-decoder structure to achieve better results. Early works [9] [10] simply treated the global feature of the input image as the visual signal, and ignored the semantic associations between salient image regions and generated words, thus making the procedure hard to predict correct words. Attention mechanisms relieve this limitation to some extent by automatically focusing the corresponding image regions

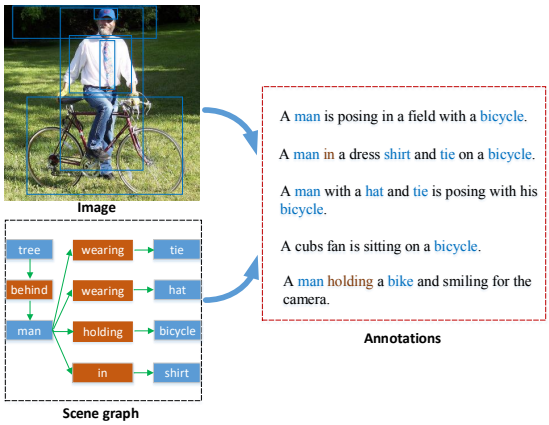

Fig. 1. The captioning model needs to infer the objects, attributes and interactions among different objects in a given image. On the one hand, the content of an image can be represented by its corresponding scene graph. On the other hand, human sentences contain important factors in the scene graph, especially the visual relationships. Therefore, we can use the scene graph to assist the generation of captions.

when producing different descriptive words, and leading to better interpretability of the model. Another major deficiency is the existence of exposure bias [11]. One of the most effective solutions for this dilemma is to incorporate reinforcement learning [7] [8], in which the proposed self-critical training [7] becomes the most reliable choice. Given a reward function, the reinforcement learning schema computes two rewards conditioned on the ground-truth sentence and the intermediate generated sentence, and the difference between these two rewards will be served as the guidance signal to correctly update the gradient direction of the model. Such strategy consequently resolves the discrepancy between training and testing, and significantly improves the quantitative and qualitative results.

Regardless of the general combination of the aforementioned strategies, a common issue not fully explored is how to cope with relationships among different visual objects within an image and eventually make the caption generation more suitable for human perception. It is obvious that

* Zhixin Li is the corresponding author (lizx@gxnu.edu.cn).

modeling visual relationships is crucial to a better understanding of the specific scenes in reality. In the domain of image captioning, visual relationships are usually included in the relevant descriptions of each image, and we can in turn use these visual relationships to guide the generation of sentences, as depicted in Figure 1. But how to efficiently combine the relation-aware contexts is still largely underexplored. In this paper we use the scene graph to represent the sophisticated visual relationships within the image, aiming to provide sufficient relationship information for boosting sentence generation. The scene graph can be viewed as a collection of visual relationships, in which each relationship pair $< subject - predicate - object >$ explicitly characterize the association between different two objects detected in an image. Note that, we view the complete representation $< subject - predicate - object >$ as an available visual relationship rather than the interaction (*predicate*) between each pair of objects. In order to obtain the graph-structured representations of the images, we train a scene graph generation model similar to previous work [12]. And we introduce a GCN based relation-aware design into our end-to-end model to incorporate relationship knowledge while attending to specific regions of the image. In particular, we build our model upon the popular Transformer architecture, leading to a significant improvement on captioning performance.

Our contributions can be summarized as followings:

- To the best of our knowledge, we are the first to introduce scene graphs into the Transformer encoder-decoder architecture for caption generation.
- We explore to convert the graph-structured semantic representation of the image into some relation-aware representations using a graph convolutional network.
- We perform comprehensive experiments on the MS COCO benchmarks, the results demonstrate that our model yields significantly better performance compared to other state-of-the-art approaches.

2. PROPOSED METHOD

In this work, we propose SGT which combines Scene Graphs and Transformer for image captioning, an overview of the proposed architecture is illustrated in Figure 2.

2.1. Scene Graph Generation

Instead of directly feeding the visual information into a captioner to generate sentence word by word, we explore to transform the image into a kind of graph structure representation referred as scene graph, to enable relation-aware modeling. The scene graph representation characterizes the objects and relationships within an image, which is typically expressed as a tuple $G = (N, E)$ where N and E represent the sets of nodes and edges, respectively.

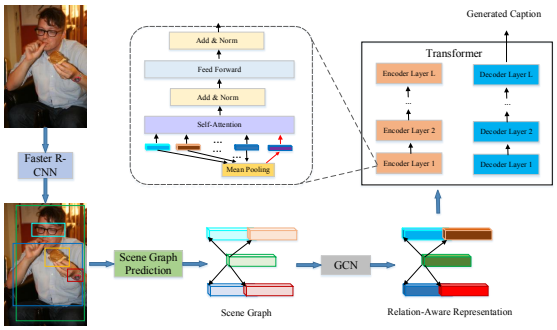

Fig. 2. Our SGT model for image captioning consists of four key components: 1) a Faster R-CNN for parsing the image into a set of object-related regions, 2) a scene graph prediction model for generating the corresponding scene graph for each image using the detected image regions, 3) a graph convolutional network for encoding the graph nodes of the scene graph into some relation-aware representations, and 4) a Transformer model that consumes the visual contexts and word contexts for inferring a syntactically and semantically correct caption. In particular, we feed the relation-aware representations into the first encoder layer. Note that the bounding box colors are consistent with the feature colors.

Notably, we follow the neural motifs model [13] to formulate our scene graph generation model. The core of the model is to use the features extracted from Faster R-CNN to train a relationship detector for identifying the relationship between each pair of objects. In our design, we remain the overall structure unchanged, but simply replace the VGG with ResNet-101 in the object detection stage so as to capture more useful visual information and improve detection performance.

2.2. Graph Convolutional Network

Now we introduce the pipeline $G \rightarrow \hat{V}$. Given a scene graph $G = (N, E)$, where N and E contain a set of object regions and relationships respectively. Since each region $\boldsymbol{v}_i$ in $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K\}$ is labeled by a specific category o_i in $O = \{o_1, o_2, \ldots, o_K\}$, each edge in the scene graph can also be denoted as a triple $\langle o_i, r_{ij}, o_j \rangle$, where r_{ij} is the relationship between o_i and o_j. To ensure the possibility of propagating information along edges of the graph, we first project the features of all objects and edges to a common space $\mathbb{R}^{D_{in}}$ with dimension of D_{in} using a learned embedding layer, thus we get input vectors $\boldsymbol{v}_i, \boldsymbol{v}_r \in \mathbb{R}^{D_{in}}$ for all nodes and edges. We then perform a single convolution network on these input vectors to compute output vectors $\boldsymbol{v}'_i, \boldsymbol{v}'_r \in \mathbb{R}^{D_{out}}$ of dimension D_{out} for all embeddings of nodes and edges. Specifically, three functions g_s, g_p, and g_o are used to convert the original triple of vectors $\langle \boldsymbol{v}_i, \boldsymbol{v}_{ij}, \boldsymbol{v}_j \rangle$ into new triple of vectors $\langle \boldsymbol{v}'_i, \boldsymbol{v}'_{ij}, \boldsymbol{v}'_j \rangle$ for the subject o_i, predicate r_{ij}, and object o_j respectively. In the factual implement, the output vectors $\boldsymbol{v}'_r$

for edges are simply computed by $\boldsymbol{v}_r' = g_p(\boldsymbol{v}_i, \boldsymbol{v}_{ij}, \boldsymbol{v}_j)$.

Since each object o_i in the graph may connect to other objects through two different types of edges, i.e., edges starting at o_i and edges terminating at o_i, computing new vector $\boldsymbol{v}_i'$ for an object o_i takes more complex steps. In the case of edges starting at o_i, we first employ function g_s to compute a candidate vector for each pair of relationship $\langle o_i, r_{ij}, o_j \rangle$, then we construct a candidate set V_i^s by collecting all of these candidate vectors. Similarly, we construct another candidate set V_i^o by employing function g_o to operate on all relationship pairs $\langle o_j, r_{ji}, o_i \rangle$ for the case of edges terminating at o_i. To this end, a symmetric function h is utilized to poor all of the candidate vectors from V_i^s and V_i^o to a single vector. The operations are shown as follows:

$$V_i^s = \{g_s(\boldsymbol{v}_i, \boldsymbol{v}_{ij}, \boldsymbol{v}_j)\} \tag{1}$$

$$V_i^o = \{g_o(\boldsymbol{v}_j, \boldsymbol{v}_{ji}, \boldsymbol{v}_i)\} \tag{2}$$

$$\boldsymbol{v}_i' = h(V_i^s \cup V_i^o) \tag{3}$$

After processing the scene graph via a series of graph convolutional layer, all of the region representations $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K\}$ have been transformed into a set of relation-aware representations $V' = \{\boldsymbol{v}_1', \boldsymbol{v}_2', \ldots, \boldsymbol{v}_K'\}$.

2.3. Relation-Aware Transformer

Instead of feeding the K representations into the first encoder layer of the transformer architecture, we augment the input set by adding the mean pooling feature $\boldsymbol{v}_0' = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{v}_i$. The $\boldsymbol{v}_0'$ can be regarded as the global context of these relation-aware representations. Thus, we get the new inputs $V' = \{\boldsymbol{v}_0', \boldsymbol{v}_1', \ldots, \boldsymbol{v}_K'\}$.

Our image captioner is built upon the recent Meshed-Memory Transformer (M^2 Transformer) model [14]. We convert the relation-aware representations V' into vectors of dimension d_k, which can be further transformed to a matrix $X \in \mathbb{R}^{N \times d_k}$. The self-attention with "memory slots" can be calculated as:

$$S_{mem}(X) = Attention(X W_Q, K, V) \tag{4}$$

$$K = [X W_K, M_K] \tag{5}$$

$$V = [X W_V, M_V] \tag{6}$$

where $W_Q \in \mathbb{R}^{d_k \times d}$, $W_K \in \mathbb{R}^{d_k \times d}$, $W_V \in \mathbb{R}^{d_k \times d}$, $M_K \in \mathbb{R}^{N \times d}$and $M_V \in \mathbb{R}^{N \times d}$ are learnable matrices, and $[,]$ denotes concatenation.

Figure 3 illustrates the meshed connectivity of the M^2 Transformer, we can see that each encoding layer is connected to all decoding layers, such scheme can make better use of different level contributions of the encoded features to predict captions. Given a sentence $Y = \{Y_1, Y_2, ..., Y_N\}$ and the outputs $\tilde{X} = \left\{\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_L\right\}$ from all encoder layers, a

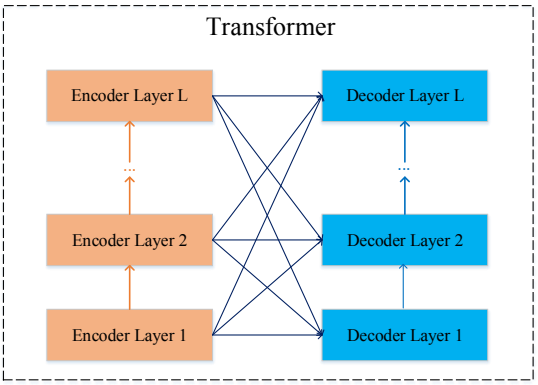

Fig. 3. The meshed connectivity between encoder layers and decoder layers in Transformer.

gated unit is added to each encoder-decoder cross-attention to measure the contributions of low- and high-level encoded features, which is calculated as:

$$S_{mesh}\left(\tilde{X}, Y\right) = \sum_{i}^{L} \alpha_i \odot C\left(\tilde{X}_i, Y\right) \tag{7}$$

$$C\left(\tilde{X}_i, Y\right) = Attention\left(Y W_Q, \tilde{X}_i W_K, \tilde{X}_i W_V\right) \tag{8}$$

$$\alpha_i = \sigma\left(\left[Y, C\left(\tilde{X}_i, Y\right)\right] W_i + b_i\right) \tag{9}$$

where $\odot$ is an element-wise product, σ represents the logistic sigmoid activation, W_i and b_i are weight matrix and bias. Note that the α_i plays the role of a unique gate that measures the contributions of each encoding layer.

To this end, we achieve our relation-aware modeling by injecting the relationship information into the modified Transformer, making the caption generation process more explainable and consistent with human perception.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

To evaluate the performance of our SGT captioning model, we conduct experiments on MS COCO dataset. MS COCO is the largest image captioning benchmark containing 164,062 images, each of which is associated with at least 5 captions. We follow the widely used Karpathy split, resulting in 113,287 images, 5,000 images and 5,000 images for training, validation and testing respectively.

In the experiments, we use the standard evaluation protocol to evaluate the performance of image captioning. We report our results on the widely adopted evaluation metrics, including BLEU [21], METEOR [22], ROUGE-L [23], and CIDEr-D [24].

Table 1. Image captioning performance on the MS COCO "Karpathy" test split.

	Cross Entropy Loss					CIDEr-D Optimization				
Methods	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D
SCST[7]	-	30.0	25.9	53.4	99.4	-	34.2	26.7	55.7	114.0
LSTM-A[15]	75.4	35.2	26.9	55.8	108.8	78.6	35.5	27.3	56.8	118.3
Up-Down[5]	77.2	36.2	27.0	56.4	113.5	79.8	36.3	27.7	56.9	120.1
HAN[16]	77.2	36.2	27.5	56.6	114.8	**80.9**	37.6	27.8	58.1	121.7
GCN-LSTM[17]	77.3	36.8	27.9	57.0	116.3	80.5	38.2	28.5	58.3	126.6
Up-Down-HIP[18]	-	37.0	28.1	57.1	116.6	-	38.2	28.4	58.3	127.2
SGAE[12]	**77.6**	36.9	27.7	57.2	116.7	80.8	38.4	28.4	**58.6**	127.8
WA-KG[19]	-	-	-	-	-	79.3	37.3	27.3	57.4	121.2
TDA-GLD[20]	-	-	-	-	-	78.8	36.1	27.8	57.1	121.1
Our-M^2Transformer	76.8	36.2	28.0	56.8	115.8	80.5	38.0	28.4	58.0	128.2
SGT	77.3	**37.0**	**28.3**	**57.4**	**118.1**	80.7	**38.7**	**28.9**	58.5	**129.9**

Table 2. Quantitative analysis with our re-implemented baseline using different beam sizes.

Beam Size	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D
1	Our-M^2 Transformer	79.1	62.9	48.0	35.9	28.1	57.1	123.7
	SGT	79.6	63.8	48.9	36,4	28.3	57.8	125.3
3	Our-M^2 Transformer	80.3	64.5	49.9	37.8	28.2	57.8	128.0
	SGT	80.4	65.0	50.3	38.3	28.9	58.3	128.0
5	Our-M^2 Transformer	80.5	64.8	50.1	38.0	28.4	58.0	128.2
	SGT	**80.7**	**65.2**	**50.7**	**38.7**	**28.9**	**58.5**	**129.9**

3.2. Comparative Analysis

3.2.1. Results on MS COCO Dataset

We first report the results on the MS COCO benchmark. As most previous works, the comparison results are reported for captioning models trained with cross entropy loss and optimized with CIDEr-D score. For fair comparison, we re-implement the M^2 Transformer model[14]. This baseline is used for verifying the importance of the specific design of incorporating scene graph structure into the sequence learning based captioning framework.

As shown in Table 1, our SGT model exhibits a promising improvement. Particularly, the BLEU-4 and CIDEr-D scores are boosted up to 37.0% and 118.1% when trained with cross entropy objective, making the improvement over the baseline by 0.8% and 2.3%, respectively. Even though optimized with CIDEr-D score, our SGT still improves the BLEU-4 and CIDEr-D scores from 38.0% to 38.7% and 128.2% to 129.9%, respectively. While compared with other state-of-the-art methods, our method exhibits better performances on almost all of the standard metrics. This demonstrates the effectiveness of the proposed relation-aware modeling.

3.2.2. Selection of Beam Size

We present additional ablation study on the MS COCO test set by applying different beam sizes. Table 2 shows the results of our re-implemented M^2 Transformer baseline and the proposed SGT model while using beam sizes of 1, 3, and 5. We observe that even using a beam size of 1, these two captioning models still achieve competitive performance, while our SGT surpasses the baseline in all evaluation metrics. In contrast, choosing a beam size of 3 makes our CIDEr-D score close to the baseline, while the best improvement is on the METEOR score, the 0.7% relative gains. As expected, using a beam size of 5 can bring remarkable performance gains on all metrics compared to using a beam size of 1. Thereby, we set the beam size as 5 in other experiments.

4. CONCLUSIONS

In this paper, we present SGT that combines Scene Graphs and Transformer. The scene graph structure establishes potential associations between each pair of object nodes, which often consists of abundant semantic contents of an image. Particularly, a graph convolutional network is employed to enrich such graph-structured representation, bringing more benefits for integrating object and relationship knowledge into some relation-aware representations. After that, the Transformer language model is incorporated to decode these relation-aware representations into a fluent sentence.

5. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China(Nos. 61966004, 61866004), Guangxi Natural Science Foundation(No. 2019GXNSFDA245018), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

6. REFERENCES

[1] Haiyang Wei, Zhixin Li, Canlong Zhang, and Huifang Ma, "The synergy of double attention: Combine sentence-level and word-level attention for image captioning," *Computer Vision and Image Understanding*, vol. 201, pp. 103068, 2020.

[2] Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi, "Integrating scene semantic knowledge into image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 2, pp. 1–22, 2021.

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image captioning with semantic attention," in *CVPR*, 2016, pp. 4651–4659.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, pp. 6077–6086.

[6] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017, pp. 375–383.

[7] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017, pp. 7008–7024.

[8] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li, "Deep reinforcement learning-based image captioning with embedding reward," in *CVPR*, 2017, pp. 290–298.

[9] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *ICCV*, 2015, pp. 2407–2415.

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[11] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[12] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019, pp. 10685–10694.

[13] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018, pp. 5831–5840.

[14] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020, pp. 10578–10587.

[15] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei, "Boosting image captioning with attributes," in *ICCV*, 2017, pp. 4894–4902.

[16] Weixuan Wang, Zhihong Chen, and Haifeng Hu, "Hierarchical attention network for image captioning," in *AAAI*, 2019, vol. 33, pp. 8957–8964.

[17] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018, pp. 684–699.

[18] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Hierarchy parsing for image captioning," in *ICCV*, 2019, pp. 2621–2629.

[19] Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma, "Boost image captioning with knowledge reasoning," *Machine Learning*, vol. 109, no. 12, pp. 2313–2332, 2020.

[20] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Transactions on Multimedia*, 2020.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*. Association for Computational Linguistics, 2002, pp. 311–318.

[22] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[23] Chin-Yew Lin and Eduard Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *NAACL*, 2003, pp. 150–157.

[24] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.