SGT: Scene Graph-Guided Transformer for Surgical Report Generation

Chen Lin[1,2], Shuai Zheng[1,2], Zhizhe Liu[1,2], Youru Li[1,2], Zhenfeng Zhu[1,2(✉)], and Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
zhfzhu@bjtu.edu.cn

Abstract. The robotic surgical report reflects the operations during surgery and relates to the subsequent treatment. Therefore, it is especially important to generate accurate surgical reports. Given that there are numerous interactions between instruments and tissue in the surgical scene, we propose a **S**cene **G**raph-guided **T**ransformer (SGT) to solve the issue of surgical report generation. The model is based on the structure of transformer to understand the complex interactions between tissue and the instruments from both global and local perspectives. On the one hand, we propose a relation driven attention to facilitate the comprehensive description of the interaction in a generated report via sampling of numerous interactive relationships to form a diverse and representative augmented memory. On the other hand, to characterize the specific interactions in each surgical image, a simple yet ingenious approach is proposed for homogenizing the input heterogeneous scene graph, which plays an effective role in modeling the local interactions by injecting the graph-induced attention into the encoder. The dataset from clinical nephrectomy is utilized for performance evaluation and the experimental results show that our SGT model can significantly improve the quality of the generated surgical medical report, far exceeding the other state-of-the-art methods. The code is public available at: https://github.com/ccccchenllll/SGT_master.

Keywords: Surgical report generation · Transformer · Scene graph.

1 Introduction

Deep learning has been widely used in computer-aided diagnosis (CAD) for the past few years [15, 30, 29]. Thereinto, computer assisted surgery (CAS) expands the concept of general surgery, which uses computer technology for surgical planning and to guide or perform surgical interventions. With the advent of CAS, general surgery has made great strides in minimally invasive approaches. For example, in the field of urology, surgical robots perform pyeloplasty or nephrectomy for laparoscopic surgery. The surgical reports are required to record the surgical procedure performed by the microsurgical robot. Automatic generation

of surgical reports frees surgeons and nurses from the tedious task of report recording, allowing them to focus more on patients' conditions and providing a detailed reference for post-operative interventions.

Surgical report generation, also called image caption, involves the understanding of surgical scenes and the corresponding text generation. In natural scenes, image caption algorithms have achieved good performance on MS-COCO [14], flicker30k [28] and Visual Genome [11], which have evolved from earlier approaches based on template stuffing [27, 1] and description retrieval [17, 6] to current deep learning-based generative approaches [22, 25]. However, in the biomedical field, studies on image caption are relatively rare. Existing studies on medical image caption more focus on radiological images, such as chest X-rays [9, 23]. However, with the spread of microsurgical robots, surgical reports generation is supposed to receive more attention. Hence, in our work, we focus on understanding surgical scene and generating accurate descriptions.

Many report generation methods follow the architecture of CNN-LSTM. Despite their widespread adoption, LSTM-based models are still affected by their sequential nature and limited representational power. To tackle this shortcoming, a fully-attentive paradigm named Transformer has been proposed [20] and has made great success in machine translation tasks [5]. Similarly, The Transformer-based model has also been applied to the report generation task. Xiong *et al.* [24] proposed hierarchical neural network architecture – Reinforced Transformer to generate coherent informative medical imaging report. Hou *et al.* [7] developed a Transformer-based CNN-Encoder to RNN-Decoder architecture for generating chest radiograph reports. Considering the excellent performance of Transformer in terms of report generation, it is also adopted as main architecture in our work.

Different from the above report generation tasks, the various instruments and complex interactive relationships in the surgical scene make it difficult to generate surgical reports. To generate accurate reports, it is necessary to understand the interactive relations in the surgical scene. However, the previous works mentioned above lack considerations of modeling the inherent interactive relations between objects. To address this issue, we propose a scene graph-guided transformer model. Unlike the previous transformer-based model, we exploit the inputs and interactive relationships in both global and local ways. In summary, the following contributions can be highlighted:

- We propose a novel surgical report generation model via scene graph-guided transformer (SGT), in which the visual interactive relationships between tissues and instruments are well exploited from both global and local ways.
- To reinforce the description of interactions in the generated report, a global relation driven attention is proposed. It uses the sampled interactive relationships with diversity as augmented memory, instead of the traditional way of utilizing the inputs directly.
- To characterize the interactive relationships in each specific surgical image, we also propose a simple yet ingenious approach to homogenize the given heterogeneous scene graph, by which a graph-induced attention is injected into the encoder to encode the above local interactions.

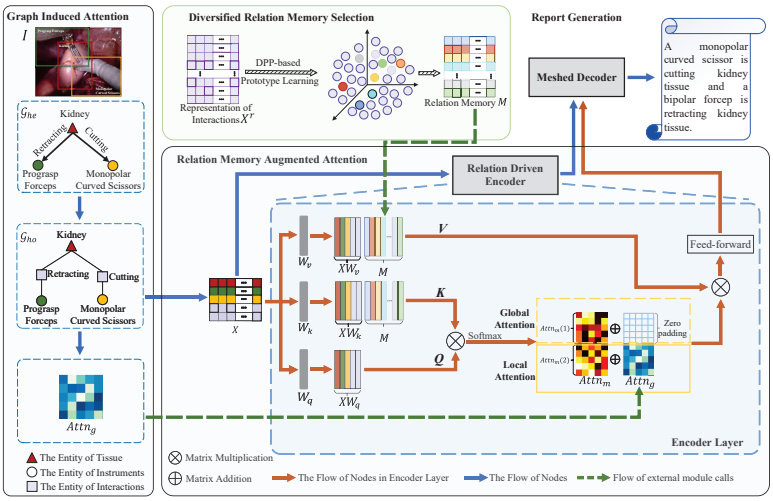

Fig. 1. The framework of the proposed SGT for surgical report generation.

2 Methodology

2.1 Overview of the Proposed Framework

For each surgical image I, a pre-built scene graph $\mathcal{G}_{he}(\mathcal{V}_{he}, \mathcal{E}_{he}, X^v_{he}, X^r_{he})$ is assumed to be available by [8], where $\mathcal{V}_{he} = \{v^i_{he}\}_{i=1,\ldots,|\mathcal{V}_{he}|}$ and $\mathcal{E}_{he} = \{e_{he}(i,j)\}_{i,j\in\mathcal{V}_{he}}$ are the sets of nodes and edges, respectively. $X^v_{he} \in \mathbb{R}^{|\mathcal{V}_{he}|\times d}$ and $X^r_{he} \in \mathbb{R}^{|\mathcal{E}_{he}|\times d}$ denote the associated representations of graph nodes and the interactive relationships between nodes, respectively. Here, the nodes refer to the visual objects extracted from image I. In particular, for a real surgical scene, $\mathcal{V}_{he}$ can be classified into two types of nodes according to their intention roles: tissue node $\boldsymbol{t}$ and instrument node $\boldsymbol{o}$, i.e., we have $v^i_{he} \in \{\boldsymbol{t}, \boldsymbol{o}\}$ for $i = 1, \ldots, |\mathcal{V}_{he}|$. Here, it is worth noting that the graph $\mathcal{G}_{he}$ can be considered a heterogeneous graph, because the links between nodes are also depicted in a representation space, which is different from the general homogeneous graph.

The overall framework of the proposed **S**cene **G**raph-guided **T**ransformer for surgical report generation, also named by SGT, is shown in Fig.1. It is divided into two main modules: the *relation driven encoder* is responsible for encoding the input heterogeneous scene graph, and the *meshed decoder* tends to read from each encoder layer to generate report words by words. Specifically, for the encoder, it receives the injections of the relation memory augmented attention and graph induced attention, to guide the representation learning of visual scene from both the global and local perspectives.

2.2 Relation Driven Attention

Diversified Sampling of Relation Memory. For the task of caption generation, when using visual objects as input, despite the fact that self-attention

can encode pairwise relationships between regions, it fails to model the a priori knowledge of the relationships between visual objects. To address this issue, an operator called persistent memory vectors was proposed in [4]. However, this persistent memory augmented attention still relies on the input image objects to establish a priori interrelationships between image objects. Different from [4], the interactive relationships between instruments and tissue can be directly obtained from the scene graph in the form of representation using the previous work [8], which could be regarded as a very beneficial prior. Hence, to enhance the description of interactive relationships in the generated surgical report, an augmented attention using a prior interaction as memory is proposed.

Let $X^r = [X_{he}^{r,1}, \ldots, X_{he}^{r,N}] \in \mathbb{R}^{n_r \times d}$ denote the relational representation of all of collected N images, where $n_r = \sum_{p=1}^{N} |\mathcal{E}_{he}^{p}|$ represents the total number of relational representations. To establish the relation memory, a straightforward way is to use X^r for it. But doing so will pose two problems. The first is the high computational complexity, and the other one is the over-smoothing of the learned attention due to the excessive redundancy of these relational representations, which will lead inevitably to the loss of focus. For this reason, a sampling scheme with diversity was considered. In particular, the determinant point process (DPP) [16] is used to sample X^r to obtain a rich diversified prototype subset of interaction representation, so that the sampled subset could cover as much of the representation space of the interactive relationships as possible.

Given Z as the metric matrix of X^r, DPP is capble of selecting a diversified subset $X_S^r \subseteq X^r$, whose items are indexed by $S \subseteq L = \{1, \ldots, n_r\}$, by maximizing the following sampling probability $P_Z(X_S^r)$ of X_S^r:

$$P_Z(X_S^r) = \frac{det(Z_S)}{det(Z+I)} \tag{1}$$

where $\sum_{S \subseteq L} det(Z_S) = det(Z+I)$, I is the identity matrix, $Z_S \equiv [Z_{ij}]_{i,j \in S}$, and $\det(\cdot)$ denotes the determinant of a matrix. As given by Eq. 1, it can be known that any subset of L corresponds to a probability, which will result in a large search range for the prototype index subset. In order to eliminate the uncertainty of the sample set capacity in the standard DPP, Kulesza [12] proposed a variant of standard DPP with fixed subset size $|S| = k$, to realize the controllability of the sampling process. Through k-DPP, the most appropriate k interaction relations can be selected as the prototypes to serve for the relation memory $M = X_S^r \in \mathbb{R}^{k \times d}$.

Relation Memory Augmented Attention. To enhance the intervention of interactive relationships in the encoder, the relation memory M are concatenated with the transformed input X to obtain the augmented key $\boldsymbol{K} = [XW_k; M] \in \mathbb{R}^{(m+k) \times d}$ and value $\boldsymbol{V} = [XW_v; M] \in \mathbb{R}^{(m+k) \times d}$, respectively, where $W_k \in \mathbb{R}^{d \times d}$ and $W_v \in \mathbb{R}^{d \times d}$ are the corresponding projection matrix, $X = [X_{he}^{v}; X_{he}^{r}] \in \mathbb{R}^{m \times d}$ with $m = |\mathcal{V}_{he}| + |\mathcal{E}_{he}|$ is the input node representation of the homogeneous graph $\mathcal{G}_{ho}$ to be described in the following section, and $[\cdot;\cdot]$ indicates the concatenation of two matrices. Furthermore, the relation memory augmented

attention can be defined as:

$$Attn_m(\boldsymbol{Q}, \boldsymbol{K}) = softmax\left(\frac{\boldsymbol{K}\boldsymbol{Q}^{\boldsymbol{T}}}{\sqrt{d}}\right) = \begin{bmatrix} Attn_m(1) \\ Attn_m(2) \end{bmatrix} \tag{2}$$

where $\boldsymbol{Q} = XW_q \in \mathbb{R}^{m\times d}$, and $\sqrt{d}$ denotes a scaling factor relevant to d. Clearly, the relation memory augmented attention $Attn_m$ consists of two attentional block matrices $Attn_m(1)$ and $Attn_m(2)$, in which $Attn_m(1)$ is obtained by calculating the pairwise similarity according to the input X itself, and $Attn_m(2)$ carries the information from the diversified relation memory prototypes M that is globally sampled in the representation space of interactive relationships.

2.3 Graph Induced Attention

The relation memory augmented attention mentioned above establishes a global perception of various interactions by the entities, i.e., nodes in the scene graph. However, such global perception is still inadequate for the portrayal of the detailed interactions among entities in a given specific scene graph. In fact, it is also crucial to seek the local perception of a particular visual scene, as a complement to global perception, when generating surgical reports.

Homogenization of Heterogeneous Graph. For the pre-established heterogeneous scene graph $\mathcal{G}_{he}$, it reflects the unique interaction between the visual objects of the associated image. However, due to its heterogeneity, direct application of the existing graph structure could achieve the local perception to a certain extent, while it is difficult to make such use of various interactions in graphs in such a way.

To address this issue, a simple yet ingenious way is to homogenize the heterogeneous link '$\boldsymbol{t} \overset{\boldsymbol{r}}{\longleftrightarrow} \boldsymbol{o}$' to the form of '$\boldsymbol{t} \longleftrightarrow \boldsymbol{r} \longleftrightarrow \boldsymbol{o}$', where $\boldsymbol{r}$ denotes the interaction between the tissue node $\boldsymbol{t}$ and instrument node $\boldsymbol{o}$. In this way, the interaction information hidden in the links in the heterogeneous graph can also be represented as nodes in a re-build homogeneous graph. Specifically, the homogenization of heterogeneous graph can be illustrated as $\mathcal{G}_{he}(\mathcal{V}_{he}, \mathcal{E}_{he}, X^v_{he}, X^r_{he}) \rightarrow \mathcal{G}_{ho}(\mathcal{V}_{ho}, \mathcal{E}_{ho}, X^v_{ho})$, where $\mathcal{V}_{ho} = \{\mathcal{V}_{he} \cup \mathcal{V}_r\}$ with $v^i_{ho} \in \{\boldsymbol{t}, \boldsymbol{o}, \boldsymbol{r}\}$, $i = 1, \ldots, |\mathcal{V}_{ho}|$, and $\mathcal{E}_{ho} = \{e_{ho}(i,j)\}_{i,j\in\mathcal{V}_{ho}}$ are the sets of nodes and edges of homogeneous graph $\mathcal{G}_{ho}$, respectively. $\mathcal{V}_r$ is the set of nodes with each node representing a specific interactive relationship and $|\mathcal{V}_r| = |\mathcal{E}_{he}|$. $X^v_{ho} = [X^v_{he}; X^r_{he}] \in \mathbb{R}^{m\times d}$, i.e., X mentioned above, denotes the associated representation of each node with $m = |\mathcal{V}_{ho}| = |\mathcal{V}_{he}| + |\mathcal{E}_{he}|$ being the number of nodes.

Particularly, without loss of generality, taking the visual scene in graph induced attention module of Fig.1 as an example, there are two instruments, *prograsp forceps* and *monopolar curved scissors*, a *kidney* tissue, and two *rectracting* and *cutting* interactions in $\mathcal{G}_{he}$. By the homogenization of $\mathcal{G}_{he}$, the two interactions including *rectracting* and *cutting* in the new converted homogeneous graph $\mathcal{G}_{ho}$ will be treated as nodes.

Attention Based on Homogeneous Graph. Given the obtained heterogeneous graph $\mathcal{G}_{ho}$, its graph structure can be represented using the adjacency

matrix A_{ho} with $A_{ho}(i,j) = 1$, if $e_{ho}(i,j) \in \mathcal{E}_{ho}$, and $A_{ho}(i,j) = 0$, otherwise. As known to all, the essence of self-attention is to look for the intrinsic correlations between inputs. Considering the adjacency matrix A_{ho} describes the connections between various entities including the tissue, instruments, and interactions, which just felicitously reflect the correlation mentioned above, we can define the graph induced attention as $Attn_g = D^{-\frac{1}{2}} A_{ho} D^{-\frac{1}{2}}$, where D is a diagonal matrix with $D_{ii} = \sum_j A_{ho}(i,j)$. Furthermore, we will have a fused attention $Attn$ via a linear combination of the relation memory driven attention $Attn_m$ with the graph induced attention $Attn_g$:

$$Attn = \begin{bmatrix} Attn(1) \\ Attn(2) \end{bmatrix} = \left[(1-\gamma) \cdot \begin{bmatrix} Attn_m(1) \\ Attn_m(2) \end{bmatrix} + \gamma \cdot \begin{bmatrix} 0 \\ Attn_g \end{bmatrix} \right] \tag{3}$$

where γ is a trade-off coefficient. Obviously, we can find from Eq.3 that $Attn$ is also composed of two attentional block metrices $Attn(1)$ and $Attn(2)$. As far as $Attn(1)$ is concerned, it explicitly carries the global interactive information contained in relation memory via global sampling based on DPP. Different from $Attn(1)$, $Attn(2)$ can be seen as a local perception of an input surgical image to some extent since it exploits effectively the information of the specific scene graph itself from two different views respectively, i.e., the node representation in $Attn_m(2)$ and the graph structure. Finally, the node embeddings can be calculated as $\boldsymbol{H} = Attn^T \cdot \boldsymbol{V} \in \mathbb{R}^{m \times d}$.

2.4 Caption Generation

To generate the word in the report sequentially, the decoder takes both the words generated in the previous stage and the output $\boldsymbol{H}$ by the encoder as input. Just like in M^2T[4], we use the same backbone structure, i.e., meshed decoder, for caption generation. Specifically, the encoder encodes $\boldsymbol{X}$ to $\boldsymbol{H}$ via the proposed dual attention, and then the decoder reads from the output encoder and finally performs a probabilistic calculation to determine the output of the next word.

3 Experiment Results and Analysis

3.1 Dataset

The dataset comes from Robotic Instrument Segmentation Sub-Challenge of 2018 MICCAI the Endoscopic Vision Challenge [2], which consists of 14 nephrectomy record sequences. The surgical report of the frames in each sequence were annotated by an experienced surgeon in robotic surgery [26] and the scene graph of each frame is generated by [8]. Specifically, there are a total of 9 objects in the scene graphs of the dataset, including 1 tissue and 8 instruments. Besides, a total of 11 interactions exists among these surgical instruments and tissues, such as manipulating, grasping, etc. For the surgical report generation task, 11 sequences including 1124 frames with surgery report are selected as the training set, and the other 3 sequences including 394 frames with surgery report as the test set. For fairness, most interactions are presented in both the training and test sets.

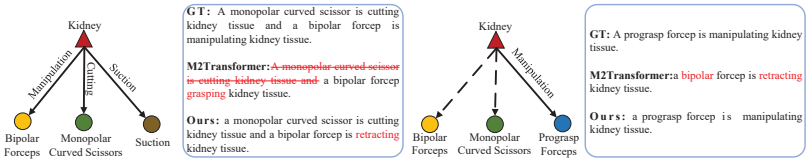

Fig. 2. Examples of the generated reports by our SGT, M^2T, and the corresponding ground-truths. $\rightarrow$ denotes interactions existing between two nodes. $\dashrightarrow$ represents there are no interactions. The words in red and the red ones with strikethrough indicate the generated incorrectly and the ground truth but are not generated, respectively.

3.2 Experimental Settings

Metrics. To quantitatively verify the effectiveness of our proposed method, the evaluation metrics used in the image caption task are applied for the surgical report generation: BLEU [19], METEOR [3], ROUGE [13], and CIDEr [21].
Implementation details. Following the preprocessing in [26], we change all the words to lowercase in each surgical report, the punctuation is removed as well. Thus, there are 45 words in the vocabulary. In our model, we set the number of heads to 8, the number of the selected prototypes k to 48, the value of γ to 0.3, and the dimensionality of the node feature d is set to 512. We train the model using cross-entropy loss and fine-tune the sequence generation using reinforcement learning following [4]. The model uses Adam [10] as the optimizer with a batch size of 10, the beam size equals to 5, and is implemented in Pytorch.

3.3 Experimental Results

Performance Comparison. To evaluate the performance of the proposed SGT, we choose to compare with several state-of-the-art transformer-based models: M^2T [4], X-LAN [18], and CIDA [26]. As shown in Table 1, our SGT significantly outperforms all the other methods in terms of all evaluation metrics. Particularly, the relative improvement achieved by SGT grows larger as the standard of BLEU becomes more stringent, indicating that the generated reports of SGT are more approximate to the real reports provided by the doctor compared to the others. Besides, to alleviate the overfitting in the case of small test set, we further conduct the experiment of 5-fold cross validation on random division of the original dataset. The results of ours (BLEU-1:0.7566, CIDEr:5.1139) are slightly lower than the original. Nevertheless, our method still outperformed other methods, the CIDEr score of ours is around 105% than M^2T. Owing to space constraints, the results of 5-fold cross validation are presented in the supplementary material. For a more straightforward comparison of SGT with other methods, some cases of the reports generated by SGT and M^2T are shown in Fig. 2. Obviously, the results of SGT are more reliable than the results of M^2T, illustrating that the proposed dual attention captures the scene information effectively.

Table 1. Performance comparisons of our SGT with other models.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
M^2T [4]	0.5881	0.5371	0.4875	0.4445	0.4691	0.6919	2.8240
X-LAN [18]	0.5733	0.5053	0.4413	0.3885	0.3484	0.5642	2.0599
CIDA [26]	0.6246	0.5624	0.5117	0.4720	0.3800	0.6294	2.8548
SGT(Ours)	**0.8030**	**0.7665**	**0.7300**	**0.6997**	**0.5359**	**0.8312**	**5.8044**
Improv.	28.56%	36.29%	42.66%	48.24%	14.24%	20.13%	103.32%

Table 2. Ablation Study of SGT.

Method		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
M	$Attn_g$							
×	×	0.7232	0.6799	0.6410	0.6104	0.5088	0.7805	5.1892
✓	×	0.7429	0.7027	0.6640	0.6343	0.5139	0.7890	5.2265
×	✓	0.7776	0.7373	0.6998	0.6716	0.5009	0.7890	5.2578
✓	✓	**0.8030**	**0.7665**	**0.7300**	**0.6997**	**0.5359**	**0.8312**	**5.8044**

Ablation Study. To fully validate the effectiveness of our proposed relation memory prototype M and graph-induced attention $Attn_g$, we conduct an ablation study to compare different variants of SGT. As shown in Table 2, M and $Attn_g$ respectively already bring an improvement in terms of the base model. It's obvious that the proposed M and $Attn_g$ bring significant performance gains. In summary, we can observe that M and $Attn_g$ are designed reasonably and the performance degrades to some extent when removing any of them. In addition, we also perform 5-fold cross validation on ablation study. The experimental results are slightly lower than the original results. But it can still be seen the effectiveness of M and $Attn_g$. The results of the 5-fold cross validation are shown in the supplementary material.

Hyper-parameter Sensitivity Analysis. We then evaluate the role of the tradeoff γ and the number of the selected prototypes. Fig. 3 intuitively shows the change trend of γ. Notably, when γ is set to 0.3, SGT achieves the best performance on all metrics. Furthermore, we report the performance of our approach when using a varying number of the selected prototypes k. As shown in Table. 3, the best results in terms of all the metrics is achieved with k set to 48.

Table 3. Sensitivity analysis of k.

Method		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
DPP	$k=6$	0.7123	0.6636	0.6255	0.5975	0.4801	0.7468	4.8093
	k=12	0.6884	0.6397	0.5957	0.5634	0.4855	0.7555	4.7155
	k=24	0.7982	0.7594	0.7247	0.6979	0.5148	0.8113	5.6977
	k=48	**0.8030**	**0.7665**	**0.7300**	**0.6997**	**0.5359**	**0.8312**	**5.8044**
	k=96	0.7776	0.7381	0.7026	0.6750	0.4951	0.7948	5.6844

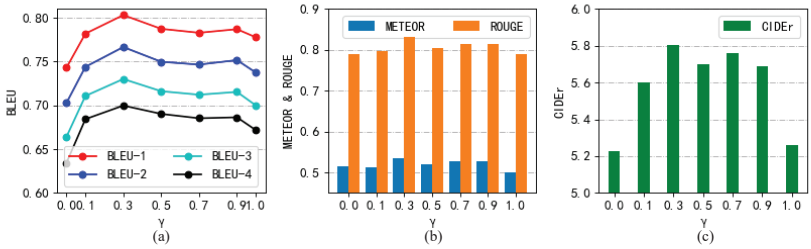

Fig. 3. Impact of the tradeoff γ on BLEU, METEOR, ROUGE and CIDEr.

4 Conclusion

In this work, we mainly focus on how to generate precise surgical reports and propose a novel scene graph-guided Transformer (SGT) model. It takes full advantage of the interactive relations between tissue and instruments from both global and local perspectives. As for the global relation driven attention, the globally sampled representative relation prototypes are utilized as augmented relation memory, thus enhancing the description of the interactions in the generated surgical report. Additionally, a graph-induced attention is proposed to characterize from a local aspect the specific interactions in each surgical image. The experiments on a clinical nephrectomy dataset demonstrate the effectiveness of our model.

Acknowledgement. This work was supported in part by Science and Technology Innovation 2030 – New Generation Artificial Intelligence Major Project under Grant 2018AAA0102100, National Natural Science Foundation of China under Grant No. 61976018, Beijing Natural Science Foundation under Grant No. 7222313.

References

1. Aker, A., Gaizauskas, R.: Generating image descriptions using dependency relational patterns. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 1250–1258 (2010)
2. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision. pp. 15–29. Springer (2010)
7. Hou, B., Kaissis, G., Summers, R.M., Kainz, B.: Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 293–303. Springer (2021)
8. Islam, M., Seenivasan, L., Ming, L.C., Ren, H.: Learning and reasoning with the graph structure representation in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 627–636. Springer (2020)
9. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
12. Kulesza, A., Taskar, B.: k-dpps: Fixed-size determinantal point processes. In: ICML (2011)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
15. Liu, Z., Zhu, Z., Zheng, S., Liu, Y., Zhou, J., Zhao, Y.: Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(2), 638–647 (2022). https://doi.org/10.1109/JBHI.2022.3140853
16. Macchi, O.: The coincidence approach to stochastic point processes. Advances in Applied Probability **7**(1), 83–122 (1975)
17. Pan, J.Y., Yang, H.J., Duygulu, P., Faloutsos, C.: Automatic image captioning. In: 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763). vol. 3, pp. 1987–1990. IEEE (2004)
18. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10971–10980 (2020)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)

22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9049–9058 (2018)
24. Xiong, Y., Du, B., Yan, P.: Reinforced transformer for medical image captioning. In: International Workshop on Machine Learning in Medical Imaging. pp. 673–680. Springer (2019)
25. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
26. Xu, M., Islam, M., Lim, C.M., Ren, H.: Class-incremental domain adaptation with smoothing and calibration for surgical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 269–278. Springer (2021)
27. Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C.: I2t: Image parsing to text description. Proceedings of the IEEE **98**(8), 1485–1508 (2010)
28. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
29. Zhang, W., Wu, H., Liu, Y., Zheng, S., Liu, Z., Li, Y., Zhao, Y., Zhu, Z.: Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis. Biomedical Signal Processing and Control **68**, 102616 (2021)
30. Zheng, S., Zhu, Z., Liu, Z., Guo, Z., Liu, Y., Yang, Y., Zhao, Y.: Multi-modal graph learning for disease prediction. IEEE Transactions on Medical Imaging pp. 1–1 (2022). https://doi.org/10.1109/TMI.2022.3159264