International Journal of Information Technology Vol. 28 No. 2 2022

Disease Knowledge Driven Transformer Network for Medical Report Generation

Yiming Cao[a], Zhen Li[a], Yonghui Xu[a] and Lizhen Cui[a]

[a]Shandong University, Jinan, China
caoyiming@mail.sdu.edu.cn, {clz,qilulizhen}@sdu.edu.cn, xu.yonghui@hotmail.com

Abstract

Automatic generation for medical image reports is a promising research problem at the intersection of computing and medicine to reduce the workload of doctors. The mainstream approaches employ the encoder-decoder paradigm to automatically generate the corresponding text reports for medical images. In this paper, we propose a Disease Knowledge Driven Transformer network (DKDT) for medical image report generation. Specifically, a graph embedding module is first utilized to extract the graph-enriched features from the input image with the guidance of a disease knowledge graph. Furthermore, DKDT adopts a Transformer-based text decoder to compile the graph-enriched features into the medical reports. The experimental results demonstrate that DKDT can effectively generate reports for medical images, which can help improve the automation and homogenization of medical reports.

Keywords: Medical report generation, Transformer, Knowledge graph.

I. Introduction

The development in deep learning has significantly impacted the medical image analysis by supporting the feature extraction from medical images to assist doctors in diagnosis. Medical reports are typically utilized in clinical as a crucial basis for developing the diagnosis. Doctors usually

describe the medical observations reflected on images by a handwritten report. However, manually writing reports for massive medical images tremendously increases the workload of doctors. In addition, the inconsistent levels of doctors and radiologists in different levels of medical institutions result in uneven quality of medical reports. Therefore, automatic generation of corresponding reports for medical images can help alleviate these problems. The existing works usually use retrieval-based methods [7] and generation-based methods [4] for automatic report generation. Some works equip pre-defined knowledge graphs [5, 15] to guide the report generation.

In this paper, we propose a Disease Knowledge Driven Transformer network (DKDT) to generate textual reports. Specifically, a CNN module is employed to extract the visual feature from input images. DKDT utilizes the graph embedding module to propagate visual features in the pre-constructed disease knowledge graph to obtain the graph-enriched features. In the disease knowledge graph, a node represents a specific disease, and related diseases are interconnected. Furthermore, the Transformer-based text decoder receives graph-enriched features and generate a coherent medical report. We experimentally evaluate DKDT with report generation methods on a real world dataset. The results show that our DKDT can generate accurate medical reports.

II. Related Work

The existing works [12, 11] usually follow the Encoder-Decoder architecture to generate reports for images. Typically, the Convolutional Neural Networks (CNN) are used as the decoder, and Recurrent Neural Networks (RNN) and self-attention networks are used as decoders. Many models [14, 10] are also furnished with attention mechanisms for the Encoder-Decoder architecture to to enhance the interpretability. Furthermore, some methods [4, 6] learn the latent representations, e.g., context vector and topic vector from the additional information by the generative models and

reinforcement learning models. Many graph-based works [13, 7, 9] organized the predefined medical knowledge into the graph form and updated the graph architectures by the graph neural network algorithms. In addition, the Transformer-based approaches [2, 8, 1] are proposed to improve the accuracy of report generation.

III. Models

The disease knowledge driven transformer network is proposed to generate consistent reports for medical images. The CNN module and graph embedding module are devised to extract the graph-enriched features from the input image. Furthermore, DKDT adopts a Transformer-based text decoder to propagate graph-enriched features in the disease knowledge graph and generate the corresponding medical image reports for the image. The overview of our proposed DKDT is displayed in Fig. 1.

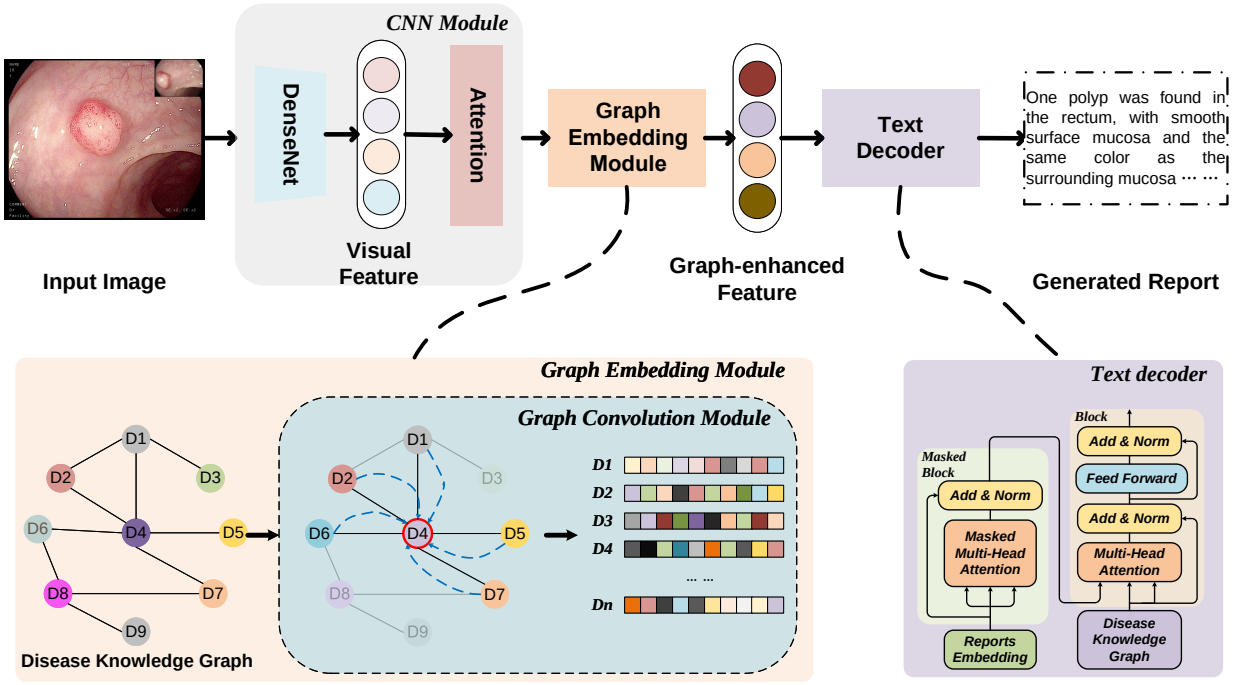

Fig 1 Overview of DKDT. The CNN module in DKDT extracts the visual features from the input image. Next, the graph embedding module is employed to obtain the graph-enriched features using the graph convolution module on the disease knowledge graph. Furthermore, the text decoder compile the graph-enriched features into textual reports.

A. CNN Module

DKDT adopts a CNN module to extract the visual features from the input image. A CNN model is first pre-trained on the datasets for the disease classification task, and then extracts features from the last convolutional layer. Subsequently, an attention mechanism is employed to feed the visual features for the next module.

B. Graph Embedding Module

The graph embedding module is used to transform visual features into graph-enriched disease features. Inspired by the work [1], we used the same disease knowledge graph to guide the transformation. Subsequently, DKDT uses the graph convolution network to propagate the visual feature on the disease knowledge graph. The graph-enriched features are output from the graph embedding module, which is the input of the text decoder.

C. Text Decoder

The Transformer-based text decoder is devised to compile the graph-enriched features into the linguistic features and generate the final reports. A masked multi-head attention is first applied to the token embeddings of medical reports. Next, the output of the masked multi-head attention is transformed into an vector Q as an input to the multi-head attention block. The other set of attention vectors K and V come from graph-enriched features $\mathbf{f}^g$. Through the feed-forward layer,

the text decoder generates the distribution D of textual output, which can be formalized as:

$$\begin{aligned} \mathbf{h}^0 &= \mathbf{X}_t\mathbf{W}_e + \mathbf{X}_p\mathbf{W}_p \\ \mathbf{h}^{l+1} &= \mathbf{block}(\mathbf{h}^l, \mathbf{f}^g) \\ D &= \text{Softmax}(\mathbf{h}^n{\mathbf{W}_e}^T) \end{aligned} \tag{1}$$

where $\mathbf{X}_t$ and $\mathbf{X}_p$ are the index of input sentences' tokens and position, $\mathbf{W}_e$ and $\mathbf{W}_p$ denote the word and position embedding matrix, $\mathbf{block}$ is the Transformer architecture.

IV. Experiment

We conducted experiments on **Gastrointestinal Endoscope image dataset (GE)** [1], which contains 15345 images and 3069 Chinese reports from the Department of Gastroenterology. We split the dataset into 7:1:2 training:validation:testing data to train and evaluate our approach.

The CNN model in the CNN module is a pre-trained DenseNet-121. The size of the input images is 512×512. the Chinese word segmentation module of Jieba [3] is used to pre-process the reports. The text decoder consists of 3 blocks with 8 attention heads. We set 512 as the dimension of all hidden states.

A. Qualitative Analysis

Fig. 2 shows visualization results of DKDT on GE. DKDT correctly predicted the rectal polyp, and the attention map also marked the position of polyps. In addition, the reports generated by DKDT is generally consistent with the reports written by doctors. For example, the report generated by DKDT covers the description for the symptoms of rectal polyps. Moreover, for the normal

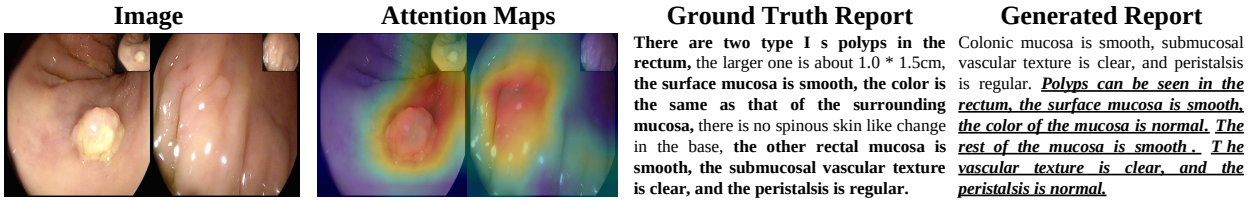

Fig 2 Sample cases of DKDT on GE. **Bold text** indicates consistency between the generated reports and ground truth. Underlined text indicates the correspondence between the generated reports and the attention maps.

findings, DKDT also give the relevant descriptions with the ground-truth normal descriptions. The visulization results demonstrate that DKDT is capable of generating reports for medical images.

V. Conclusion

In this paper, we propose a disease knowledge driven Transformer network to generate medical reports. We design a graph embedding module to extract graph-enriched features. The text decoder is adopted to generate textual reports. Experiments on a real dataset show the effectiveness of our DKDT. For future work, it is more challenging to combine knowledge and hierarchy structures to interpret the generated reports.

VI. Acknowledgments

This work is partially supported by the NSFC No.91846205; National Key R&D Program of China No. 2021YFF0900800; Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (No. 2021CXGC010506 and NO.2021CXGC010108); the State Scholarship Fund by the China Scholarship Council (CSC).

References

[1] Yiming Cao, Lizhen Cui, Fuqiang Yu, Lei Zhang, Zhen Li, Ning Liu, and Yonghui Xu. Kdtnet: Medical image report generation via knowledge-driven transformer. In *DASFAA 2022*, pages 117–132, 2022.

[2] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *ACL 2021*, pages 5904–5914, 2021.

[3] Jieba. "jieba" chinese text segmentation: built to be the best python chinese word segmentation module. https://github.com/fxsjy/jieba, 2018.

[4] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *ACL*, pages 2577–2586, 2018.

[5] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, pages 6666–6673, 2019.

[6] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*, pages 1537–1547, 2018.

[7] Xiaodan Liang, Zhiting Hu, Hao Zhang, Liang Lin, and Eric P. Xing. Symbolic graph reasoning meets convolutions. In *NeurIPS*, pages 1858–1868, 2018.

[8] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE CVPR 2021*, pages 13753–13762, 2021.

[9] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS 2018*, pages 8344–8353, 2018.

[10] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-

critical sequence training for image captioning. In *IEEE CVPR*, pages 1179–1195, 2017.

[11] Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, pages 2497–2506, 2016.

[12] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[13] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, volume 11218, pages 711–727, 2018.

[14] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE CVPR*, pages 4651–4659, 2016.

[15] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *AAAI*, pages 12910–12917, 2020.