CID at RRG24: Attempting in a Conditionally Initiated Decoding of Radiology Report Generation with Clinical Entities

Yuxiang Liao*, **Yuanbang Liang***, **Yipeng Qin, Hantao Liu, Irena Spasić**
School of Computer Science and Informatics, Cardiff University, UK
{liaoy11, liangy32, qiny16, liuh35, spasici}@cardiff.ac.uk

Abstract

Radiology Report Generation (RRG) seeks to leverage deep learning techniques to automate the reporting process of radiologists. Current methods are typically modelling RRG as an image-to-text generation task that takes X-ray images as input and generates textual reports describing the corresponding clinical observations. However, the wording of the same clinical observation could have been influenced by the expression preference of radiologists. Nevertheless, such variability can be mitigated by normalizing textual reports into structured representations such as a graph structure. In this study, we attempt a novel paradigm for incorporating graph structural data into the RRG model. Our approach involves predicting graph labels based on visual features and subsequently initiating the decoding process through a template injection conditioned on the predicted labels. We trained and evaluated our model on the BioNLP 2024 Shared Task on Large-Scale Radiology Report Generation and submitted our results to the ViLMedic RRG leaderboard. Although our model showed a moderate ranking on the leaderboard, the results provide preliminary evidence for the feasibility of this new paradigm, warranting further exploration and refinement.

1 Introduction

Radiology Report Generation (RRG) seeks to liberate radiologists from the repetitive reporting process, allowing them to focus on revising the reports and thereby enhancing the accuracy and efficiency of clinical communication. As a multi-modality task, RRG models usually employ the encoder-decoder architecture, where the encoder is a vision model that is responsible for extracting visual features from radiology images while the decoder is a language model that is responsible for converting visual features into narrative reports. Compared with the general image captioning task, the clinical observations in radiology images are more subtle. Moreover, the wording of the same clinical observation could have been influenced by the expression preference of radiologists. This raises a challenge to the model's learning ability in terms of extracting fine-grained visual features and generate accurate clinical narratives.

Our recent review of this field proposed that structured reports can alleviate the inherent diversity of natural language, thus contributing to more accurate results in the model training and evaluation (Liao et al., 2023). Benefiting from the advent of RadGraph (Jain et al., 2021), a graph-based representation of clinically significant fine-grained information extracted from reports, recent research has commenced utilising such structured representation of reports to enhance the RRG models. Relevant studies can be broadly classified into two paradigms. One paradigm fuses the graph features with visual features, letting the decoder learn how to generate the next word from a given input and the fused features (Wang et al., 2022; Yan et al., 2023; Yang et al., 2022; Li et al., 2023). Another paradigm focuses on graph generation based on the visual features and decouples the visual features from the decoding stage, allowing the language model to learn solely how to generate text based on the predicted graph (Nooralahzadeh et al., 2021; Xiong et al., 2024).

This has sparked our interest, as it raises a research question of whether there exists a new paradigm that can explicitly leverage graph structures to improve the quality of generative language models, while also enabling visual features to supplement the predicted graph with missing information. Based on this idea, we attempt a novel approach, whereby the predicted graphs are fed into a template prompt, replacing the traditional special token as the initial input to the decoder, aiming to enable a clearer query to the associated

*Contributed equally.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 591–596

image features during the generation process.

2 Related Work

In early research on RRG, many studies introduced disease labels to enhance their models (Jing et al., 2018; Yin et al., 2019; Yuan et al., 2019; Harzig et al., 2019; Wang et al., 2018). As research progressed, some studies began to explore the use of graph to replace disease labels as it can represent more fine-grained information (Zhang et al., 2020; Li et al., 2019). The graph is considered as a normalized representation of a report in terms of the the key information entities and their relationships (Jain et al., 2021).

To utilise graph data, Nooralahzadeh et al. (2021) and Yan et al. (2023) proposed modelling RRG as a pipeline of image-to-graph and graph-to-text tasks. Xiong et al. (2024) followed the same paradigm although their study contained only the first part. In contrast, Wang et al. (2022) interpreted the RRG as an image-to-text task where a graph prediction module was appended to the visual encoder. Additionally, the graph features were combined with visual features and passed to the text decoder, allowing the text decoder to learn to attend to different features. Yang et al. (2022) and Li et al. (2023) employed a similar feature fusion approach, yet their graph was not directly predicted from visual features, but rather retrieved from the paired report of a similar image identified by comparing their visual features.

3 Method

3.1 Vision Encoder Decoder Model

Our model comprises a pre-trained Transformer-based vision model as the encoder and a pre-trained language model as the decoder. A cross-attention layer and a language model head are appended to the decoder to support generation.

Let I denote a radiology image and T denote the corresponding report text. A cross-attention feature $\mathbf{\Phi}_{T,I}$ is computed by Attention$(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, where the query $\mathbf{Q}$ represents the encoded text features $\mathbf{\Phi}_T$, and the key $\mathbf{K}$ and value $\mathbf{V}$ represent the encoded visual features $\mathbf{\Phi}_I$. During the training stage, $\mathbf{\Phi}_{T,I}$ is passed to a language model head to generate a complete text sequence $\hat{T}$ at once. The model is updated by the cross-entropy loss between the probability of the predicted tokens in $\hat{T}$ and the target tokens in T. During the inference stage, the model takes an image as the encoder input and a special token as the decoder input and generates the next token through an auto-regressive decoding process.

In this architecture, the prevailing methods that combine the graph or label features with visual features can be interpreted as providing more information to K and V to be queried. However, we assume that the visual features have sufficient information, thus, we aim to enhance Q to better utilise the information from the visual features.

3.2 Graph Label Selection

We first customized a structured reporting tool based on RadGraph to preprocess the raw text. RadGraph is an information extraction tool that can convert narrative radiology reports into graphs. In RadGraph, each node is an entity that corresponds to a continuous span of text. Each edge is a uni-directional relation that connects two entities. Entities are assorted into four types: Anatomy, Observation-Present/Absent/Uncertain. Relations are assorted into three types: Suggestive-Of, Located-At, and Modify. We refer the reader to the original paper for details (Jain et al., 2021). We refined RadGraph by combining the Observation and Anatomy nodes that are linked with a Located-At edge such as "lung hyperinflate", while the other nodes were omitted. Label's text content was lemmatized. We selected labels that have appeared in more than 5,000 reports, resulting in 79, 22 and 10 label classes representing present, absent, and uncertain, respectively. For any other label, we assigned a dummy label to represent the corresponding category. Therefore, each report can be enhanced by 114 informative labels.

3.3 Multi-label Classification

Let the L^{ctg} denote the labels of a specific category $ctg = \{present, absent, uncertain\}$ extracted from a report T. We first introduced an auxiliary task of multi-label classification (MLC) between the encoding-decoding process:

$$\mathbf{p}^{ctg} = \sigma(\text{FFNN}(\theta^{ctg}; \overline{\mathbf{\Phi}_I})), \quad (1)$$

where FFNN represents a feed-forward neural network classifier with learning parameters θ that predicts the probability distribution $\mathbf{p}^{ctg}$ of labels in a specific category ctg, taking the average pooled visual features $\overline{\mathbf{\Phi}_I}$ as input to get optimised θ^{ctg}. The classification loss is computed by the cross-entropy loss between the predicted probabilities and the target labels for all categories.

By incorporating the MLC task into the model, the overall objective is thus to optimize the text generation loss and label classification loss, denoted as $\mathcal{L}_{all} = \lambda_T \mathcal{L}_T + \lambda_{MLC} \mathcal{L}_{MLC}$, where λ_T and λ_{MLC} are pre-defined weights that balance two losses.

3.4 Conditionally Initiated Decoding

To enhance the query Q in the cross-attention layer, we inject the labels directly into the decoder as its initial input. Specifically, the labels are rewritten as a label text sequence via a template: "Observation present: []; absent: []; uncertain: []. Describe them in detail: ". Each label string is filled into one of the brackets according to its category while the dummy label is filled as "others".

During training, we employ a teacher-forcing approach that uses the target labels as the source labels to fill the template. Therefore, the decoder input sequence is formed as "<BOS>label text sequence<EOS><EOS>report text sequence<EOS>". During the inference stage, we combine the predicted labels from the three classifiers and select no more than top-k labels with probabilities exceeding the threshold as the source label for the template. Therefore, the initial decoder input is transformed into "<BOS>label text sequence<EOS><EOS>" and the next token is generated through an autoregressive decoding process. A workflow of our model is illustrated in Figue 1.

3.5 Batch Inference

When performing batch inference on the data, the inconsistency in the number and length of the activated label poses an alignment issue when constructing the input tensor. To address this, we employ left padding during the inference stage to ensure the generated tokens and the initial decoder input are semantically continuous. Furthermore, the padding tokens are also marked out from the decoder attention mask to prevent them from influencing other tokens.

4 Experiments

4.1 Experimental Settings

Our experiments are conducted on the BioNLP 2024 Shared Task on Large-Scale Radiology Report Generation (Xu et al., 2024), which proposes the first standard to the community regarding the use of the dataset and evaluation metrics.

4.1.1 Datasets

This shared task provides the first large-scale collection of RRG datasets based on MIMIC-CXR (Johnson et al., 2019), CheXpert (Chambon et al., 2024), OpenI (Demner-Fushman et al., 2015), PadChest (Bustos et al., 2020) and CANDID-PTX (Vayá et al., 2020). Each data item represents a radiology examination consisting of at least one X-ray image and two pieces of text corresponding to the findings and impression sections of the radiology report. Any non-English reports were translated into English via GPT-4. The provided dataset has been split into training, validation, testing subsets. Testing data were further split into public and hidden subsets.

4.1.2 Metrics

The models are automatically evaluated by the ViLMedic metric package (Delbrouck et al., 2022b) using the following metrics: Bertscore (Zhang et al., 2019), Bilingual Evaluation Understudy: 4-gram (BLUE-4) (Papineni et al., 2002), Recall-Oriented Understudy for Gisting Evaluation: Longest Common Subsequence (ROUGE-L) (Lin, 2004), F1-RadGraph: partial (Delbrouck et al., 2022a) and all-micro-F1-CheXbert (Smit et al., 2020).

4.1.3 Implementation Details

Our model uses Swinv2-base (Liu et al., 2022) as the visual encoder and Roberta-base (Liu et al., 2019) as the text decoder. The encoder takes only the first image as input for each data. The decoder input sequence accepts a maximum of 512 tokens, where any surplus tokens are truncated. The decoder input sequences are padded to the longest sequence in each batch. We trained the model on the finding and impression respectively. In all experiments, the model was trained on NVIDIA RTX 4090 24G for 30 epochs using a learning rate of 1e-4 and a batch size of 12. A weight decay of 0.01 is set to the encoder and decoder. We updated the model with the AdamW optimizer using a linear scheduler with a warmup ratio of 0.1, and a gradient clipping set to 1. λ_T and λ_{MLC} are set to 1 and 5, respectively. During inference, we adopt the beam search strategy and set the beam size to 3 and the maximum generation length to 128. For the conditionally initiated decoding, we selected no more than 10 labels with probabilities exceeding 0.5.

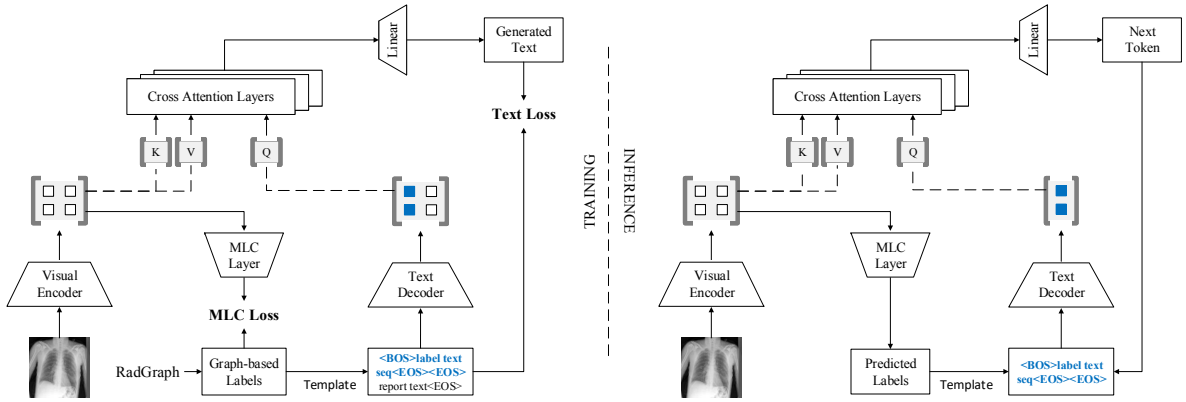

Figure 1: The workflow of our model during training (left) and inference (right). The blue text represents the conditionally initiated decoder input, which substitutes the original special token that functioned as the decoder's first input.

4.2 Results and Discussion

The performance of our model in generating the findings and impression sections of a report are illustrated in Table 1. The performance gap between the findings and impression sections is mainly due to the early termination of training to meet the system submission deadline. Although our model only exhibits a medium result on the ViLMedic RRG leaderboard (Delbrouck et al., 2022b), we assume this prototype model is feasible and has the potential to be improved.

Table 1: Model performance on the public and hidden test subsets.

Data subsets	BLEU4	ROUGEL	Bertscore	F1-cheXbert	F1-RadGraph
Public test-set					
Findings	8.29	24.38	52.28	51.13	22.26
Impression	5.25	18.71	41.72	42.86	15.13
Hidden test-set					
Findings	7.46	23.3	50.89	50.47	21.45
Impression	7.13	20.41	43.67	39.64	15.19

Firstly, we utilised only the first image from each data item as the encoder input. Given that a radiologist may refer to multiple images when composing a report, using the image features extracted from a single image may result in information loss when multiple images are available. However, the number of available images for each data item is uncertain, raising a challenge to the visual model in terms of its adaption.

Secondly, properly utilising the graph data remains unexplored yet has direct impacts on various aspects of the model. For example, the selection of graph labels can directly affect the learning difficulty of multi-label classification (MLC). If the number of labels is too small, the amount of information provided to the Conditionally Initiated Decoding (CID) may be limited even with good MLC performance. Conversely, if the number of labels is too large, the MLC performance may be significantly affected, making it impossible to provide accurate information to the CID during inference. Currently, our MLC on the finding section achieved precision/recall of 76%/40%, 63%/39% and 31%/15% on the present, absent, and uncertain labels, respectively. The trade-off between these factors requires further study. Besides, the impact of the label text template on the decoder remains unclear.

Thirdly, the current selection of model hyperparameters and the base pre-trained models for the encoder and decoder was based on experience. Due to time constraints, we did not systematically explore other combinations. Comprehensive experiments with the hyperparameters and the pre-trained models are also required in future work.

5 Conclusion

In this study, we propose a novel approach for utilizing graph structural data to support RRG. This approach involves predicting graph labels based on visual features and leveraging the predicted labels to initialize the decoder input through a template injection. We evaluated our model following the BioNLP 2024 Shared Task 1: Radiology Report Generation, where the results have been submitted to the ViLMedic RRG leaderboard. We discuss the limitations of our preliminary RRG model and the initial experiments and outline several directions for improving our model. Our model and codes are available on GitHub (Liao, 2024).

6 Limitations

Firstly, our model only accepts a single radiology image per data item as input, whereas the data item could contain multiple images, resulting in the loss of significant input information. Secondly, it remains uncertain to what extent the quality of the generated text is influenced by the decoder input initialized with graph-structured data. Thirdly, the selection of current hyperparameters and pretrained models is based on intuition rather than appropriate experimentation. More details have been discussed in Section 4.2.

Finally, our model requires an additional GPU-CPU-GPU switch during inference, leading to increased time costs. Specifically, the Conditionally Initiated Decoding process requires switching to the CPU to dynamically construct the decoder input with a tokenizer for each batch. However, we suppose that this issue can be addressed by pretokenizing and caching the template text and all graph labels. The improvement the model efficiency will be conducted in our future work.

References

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *Preprint*, arXiv:2405.19538.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. 2019. Addressing data bias problems for chest x-ray image report generation. In *British Machine Vision Conference*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.

Yuxiang Liao. 2024. Code for cid at rrg24: Attempting in a conditionally initiated decoding of radiology report generation with clinical entities. Github. Accessed on: March. 03, 2024.

Yuxiang Liao, Hantao Liu, and Irena Spasić. 2023. Deep learning approaches to automatic radiology report generation: A systematic review. *Informatics in Medicine Unlocked*, 39:101273.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022. A medical semantic-assisted transformer for radiographic report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 655–664, Cham. Springer Nature Switzerland.

Yiheng Xiong, Jingsong Liu, Kamilia Zaripova, Sahand Sharifzadeh, Matthias Keicher, and Nassir Navab. 2024. Prior-radgraphformer: Prior-knowledge-enhanced transformer for generating radiology graphs from x-rays. In *Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology: 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings*, page 54–63, Berlin, Heidelberg. Springer-Verlag.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Sixing Yan, William K. Cheung, Keith Chiu, Terence M. Tong, Ka Chun Cheung, and Simon See. 2023. Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, 42(8):2211–2222.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510.

Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 728–737.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 721–729, Cham. Springer International Publishing.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Loddon Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *AAAI Conference on Artificial Intelligence*, volume 34, Palo Alto, California USA. AAAI Press.