Knowledge-enhanced Vision-to-Language Multimodal Models for Radiology Report Generation

Yongli Mou[1][0000−0002−2064−0107]

RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany
mou@dbis.rwth-aachen.de

Abstract. Current multimodal models for radiology report generation often lack the semantic depth and contextual understanding necessary for producing clinically relevant, easy-to-read, and accurate reports. The situation is even more challenging due to the complex nature of medical imaging and the specialized language and medical terminologies in radiology reports. The gap in domain-specific knowledge in current deep learning models underscores the necessity for approaches that integrate specialized radiological expertise into advanced language models. In this research, we propose knowledge-enhanced approaches to address the problem by introducing a knowledge enhancement module in the vision-to-language multimodal model. Our research not only contributes to the field of semantic web research by demonstrating the potential of knowledge graphs in enhancing artificial intelligence models but also aims to revolutionize the radiology reporting process by automating it with greater accuracy, thereby reducing the workload of radiologists and mitigating the risk of human error.

Keywords: Radiology report generation · Multimodal learning · Knowledge graph · Semantic Web

1 Introduction

Radiology is a vital component of modern medical practice and plays a crucial role in medical diagnosis, treatment planning, and monitoring of diseases, which involves the interpretation of various medical imaging modalities such as X-rays, CT scans and MRI. From detecting fractures and tumors to monitoring the progression of chronic diseases, radiology is integral to both acute and long-term patient care. The integration of artificial intelligence (AI) in medical imaging has revolutionized the field of radiology. In recent decades, medical image analysis has achieved tremendous advancements in a variety of tasks such as classification and segmentation enhancing the ability of clinicians to interpret complex imaging data with greater accuracy and efficiency, which is primarily driven by the rapid development of deep learning [4]. Despite these technological advances, the generation of radiology reports remains a largely manual and labor-intensive process. Radiologists often face a substantial workload, and the generation of detailed accurate reports can be time-consuming. This situation is compounded by the global shortage of skilled radiologists and the variability in report quality due to human factors such as fatigue and cognitive biases [25]. Errors or inconsistencies in radiology reports can have significant implications for patient care, making the need for improvement in this area both urgent and critical. The aforementioned challenges and potential errors drive the need for accurate and automated solutions.

Automated radiology report generation (RRG) aims to generate descriptive text from a set of radiographs, which is a cross-modal text generation task. Recently, vision-language pre-training approaches such as CLIP [24] and BLIP [14, 13] have received tremendous success on various multimodal downstream tasks including image caption [26], visual question [20], etc. However, they are still subject to several significant challenges when applied in RRG, stemming from the complexity of medical imaging and the nuances of language used in radiology reports. One of the key challenges is that radiology reports are often rich in specialized complex medical terms, abbreviations, and expressions that describe the anatomy, and any abnormalities or changes observed, and sometimes suggest potential diagnoses. However, current deep learning models often lack domain-specific knowledge, which is crucial for accurate and contextually relevant radiology report generation. This gap highlights the need for a specialized approach that combines the strengths of advanced vision-language multimodal models and domain-specific knowledge in radiology.

By making data machine-readable and semantically rich, the semantic web and knowledge graphs greatly complement the domain of AI allowing for more sophisticated and context-aware algorithms [23]. In the medical domain, KGs have emerged as a powerful tool in organizing and representing complex healthcare data and knowledge. In this research, we explore the concept of Knowledge-enhanced vision-to-language multimodal models for radiology report generation aiming to leverage the synergistic capabilities of transformer-based multimodal models in processing natural language and visual inputs, along with the infusion of domain-specific knowledge from medical kGs, to automate the generation of accurate radiology reports. This research aims to contribute meaningfully to Semantic Web research, showcasing its potential to enhance the semantic richness of domain-specific knowledge for AI model training. Central to this endeavor is the utilization of medical knowledge graphs that link from various sources such as clinical data, research, drug information, etc., and provide a comprehensive understanding of domain knowledge.

2 State of the Art

Early approaches for automated RRG can be classified into two main categories, retrieval-based and disease classification-based approaches. Recent advancements in deep learning have drawn significant attention in the field of RRG. In this section, we focus on deep learning-based approaches.

Contrastive pre-training has received tremendous success in multimodal learning, especially in vision-language pretraining. Contrastive Language–Image Pre-training (CLIP) [24] involves a simple pre-training task that leverages the vast amount of text paired with images available on the internet. Utilizing loss such as InfoNCE [21], CLIP pulls image embedding and corresponding text embedding closer and pushes unpaired image and text farther in the embedding space. It can scale to achieve competitive zero-shot performance across a variety of image classification datasets. Following the idea of CLIP [24], You et al. [32] propose CXR-CLIP multi-view supervision [16] combining with image contrastive loss and text contrastive loss for enhancing the learning of study-level features of medical images and reports. Similarly, Li et al. [15] adopt BLIP [14, 13] for image-language multimodal learning, which employs a multimodal mixture of encoder-decoder architecture and utilizes the image-text contrastive loss to align the vision and language representations, image-text matching loss to learn cross-attention features between positive and negative image-text pairs, and finally language modeling loss for generating reports.

Knowledge enhancement approaches have recently emerged to address the lack of domain-specific knowledge problems of language models, as radiology reports are dense with specialized medical terms. You et al. [31] construct a joint semantic space to align visual and textual features and predict medical terms compared with gold standard labels such as MeSH-annotated labels. Similarly, Kale et al. [12] predict medical terms based on the extracted visual features and perform a disease classification upon the predicted terms. Not only medical terms (such as disease categories, organs and attributes) but also relations that usually indicate the associations among different terms (such as anatomical location, diagnostic indication, disease attributes, and health status), Dalla et al. [5] employ a triple extractor to extract the set of triples. To address the limitation of approaches based on predicted knowledge graphs such as incomplete information, Li et al. [15] utilize a pre-constructed knowledge graph to assist the report generation process, however, their approach faces knowledge noises involved during the dynamic graph construction process caused by retrieved reports.

3 Problem Statement and Contributions

3.1 Problem Statement

The overarching goal of this research is to develop a model capable of generating accurate, coherent, and clinically relevant radiology reports from radiographs. The central problem to be addressed in this research is the lack of domain-specific knowledge of current deep learning-based models causing the inefficiency and inconsistency in the process of generating radiology reports. In this research, we explore the concept of knowledge-enhanced vision-to-language multimodal models specifically designed for radiology report generation. The incorporation of advanced AI techniques, particularly those involving multimodal learning and knowledge fusion, promises to not only streamline this process but also enhance the consistency and quality of radiology reports.

Research Questions To address this problem, the research is guided by the following questions:

- RQ1: How can we extract knowledge from the visual representation of radiographs?
- RQ2: How can we model and fuse the knowledge for enhanced report generation?
- RQ3: How should we train the multimodal models that are most effective for improving multimodal learning in the context of radiology report generation?

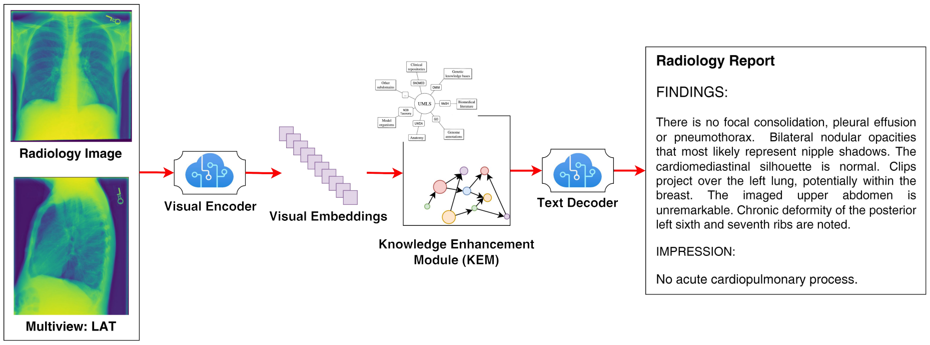
Research Hypothesis Based on these questions, the research hypothesis is formulated as follows: "*The integration of knowledge enhancement module into the transformer-based vision-to-language multimodal model will significantly improve the accuracy, consistency and clinical relevance of automated radiology report generation*". This hypothesis is based on the assumption that transformer-based crossmodal models are capable of synergizing the embeddings of visual features (radiology images) and structured knowledge (ontologies, medical lexicons, etc.) to produce more accurate, consistent, semantic richness, and clinically relevant radiology.

3.2 Contributions

A novel aspect of this research is the integration of semantic web technologies, in particular knowledge graphs, into the multimodal deep learning model, which enhances the model's capacity to utilize complex medical terminologies and relations and enrich the semantic quality of radiology reports. A significant part of this research will involve a thorough empirical evaluation of the model against traditional and existing automated methods, focusing on clinical relevance and accuracy. By applying knowledge graphs to a practical healthcare domain, this research aims to contribute meaningfully to Semantic Web research, showcasing its potential to enhance data interpretation and utility in medical contexts.

4 Research Methodology and Approach

After reviewing of relevant literature, we have identified the research gap we are going to tackle in this research, i.e., the lack of domain-specific knowledge in the current deep learning-based approaches, which also leads to the formulation of research questions to guide the research. In this section, we outline the methodology employed in developing the proposed knowledge-enhanced image-to-language multimodal model for radiology report generation. Drawing from insights garnered during the literature review, we design our architectural framework depicted in Figure 1. Central to this architecture is the incorporation of a Knowledge Enhancement Module (KEM) as shown in Figure 2, aimed at extracting and integrating relevant domain knowledge into the model.



Radiology Report

FINDINGS:

There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.

IMPRESSION:

No acute cardiopulmonary process.

Fig. 1. Architecture overview of knowledge-enhanced image-to-language multimodal model for radiology report generation.

4.1 Architecture Overview

As shown in Figure 1, we adopt an encoder-decoder architecture that is widely used in current state-of-the-art approaches [19] and introduce a Knowledge Enhancement Module (KEM) for knowledge extraction, processing and fusion. Formally, the model employs a visual encoder f_v to extract visual features $\mathcal{H}^v = f_v(\mathcal{I})$ from a radiograph $\mathcal{I}$. The KEM module takes visual embedding as input and extracts, processes and fuses the knowledge into the joint embedding $\mathcal{K} = KEM(H^v)$. Finally, a text decoder f_t translates the fused features into report texts $\hat{\mathcal{R}} = f_t(\mathcal{K})$. Deep learning-based approaches utilize radiology reports $\mathcal{R}^*$ written by professional radiologists as the reference for the corresponding radiographs and the objective is to minimize the discrepancies of the output of the model $\hat{\mathcal{R}}$ and the reference report $\mathcal{R}^*$.

4.2 Knowledge Enhancement Module

To enrich the semantics of the extracted visual features, we propose the knowledge enhancement module for extraction, processing, and fusion of domain-specific knowledge as shown in Figure 2.

Knowledge Extraction and Refinement A knowledge extraction network f_{ke} is employed to predict included knowledge $\mathcal{G}^m = f_k(H^v)$ such as medical terms and relations, then the extracted knowledge is refined and used to construct a sub knowledge graph $\mathcal{G}_{sub}$ from large scale medical knowledge graphs such as UMLS. To address RQ1, we will employ various techniques for extracting domain-specific knowledge from the visual representation of radiographs. This may include but is not limited to the utilization of attention mechanisms to identify salient regions and features within the radiographs and the incorporation of domain-specific rules and heuristics to guide the feature extraction process.

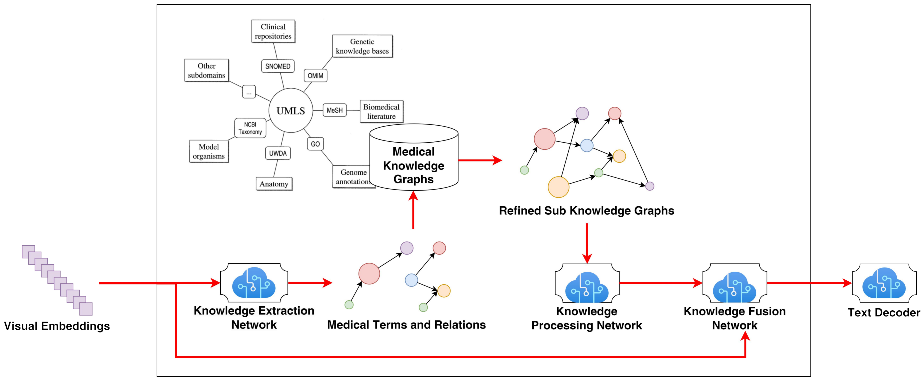


Fig. 2. Knowledge Enhancement Module (KEM) for knowledge extraction, processing and fusion.

Knowledge Processing and Fusion Addressing RQ2 involves devising effective strategies to model and fuse the extracted knowledge (in the form of graphs) for enhanced report generation. Traditional embedding methods can not well model such kind of multi-relational data. We employ the variants of graph attention network [28] from [30] for modeling and processing knowledge graphs. Similar to GreaseLM [34], we utilize a cross-modal fuser $\mathcal{K} = f_f(\mathcal{H}^v, f_g(\mathcal{G}_{sub}))$ to integrate the embeddings of visual and knowledge graph features.

4.3 Multimodal Model Training

To address RQ3, we will investigate various training strategies to optimize the effectiveness of multimodal models for radiology report generation.

We leverage recent advancements in multimodal learning and contrastive learning techniques, such as [35, 7], to learn a joint embedding across different modalities, i.e., images, graphs and text, and to bridge the modality gap, enhancing the interchangeability of embeddings.

Contrastive learning can drive a variety of pretext tasks, however, most studies follow instance discrimination tasks, which consider a query and a key as a positive pair if they originate from the same image-text pair, and otherwise as a negative sample pair. However, image-text pairs in the radiology report dataset could be correlated as the same disease causes the same symptom and observation. More effective contrastive learning approaches need to be explored.

5 Evaluation Plan

For the validation of the hypothesis and the evaluation of the effectiveness of our proposed approaches for RRG, we describe our evaluation plan in this section, which is designed to measure the accuracy, consistency, and clinical relevance of the reports generated by our proposed models, comparing them with traditional manual methods and state-of-the-art models.

5.1 Datasets

The success of any deep learning model heavily relies on the quality and quantity of the data used for training and testing. In this research, we will collect a diverse dataset of radiographs along with their corresponding expert-generated radiology reports or annotations, as well as medical knowledge graphs that models acquire domain-specific knowledge.

Radiograph Datasets Essential to our model's training and evaluation are extensive datasets comprising radiographs and corresponding reports. The datasets selected include MIMIC-CXR [10, 11], IU X-Ray [6], NIH ChestX-ray [29], Chexpert [8], and for multilingual capabilities, datasets like CX-CHR [17] and PadChest [3].

Medical Knowledge Graphs Our models not only rely on image and text data but also integrate substantial medical knowledge from various sources, such as knowledge graphs. The integration of knowledge graphs aims to enhance the clinical accuracy and relevance of the generated reports. The key knowledge graphs we are incorporating are UMLS [2] and RadGraph [9].

5.2 Metrics

Our evaluation plan employs a multifaceted set of metrics to measure the accuracy, consistency, and clinical relevance of the reports generated by our models, offering a comprehensive comparison with traditional manual methods and state-of-the-art models. In general, our chosen metrics for RRG are categorized into three types based on the target granularity, namely, entity level, graph level, and report level. The evaluation metrics are detailed as follows:

1. **Entity Recognition and Graph Construction**: Metrics such as accuracy, recall, precision, and f1-score are used to evaluate the clinical entity recognition. To ensure the clinical utility of the generated reports, we also calculate the accuracy of the diagnoses provided in the generated reports compared to the diagnoses concluded from the reference standard. We evaluate graph construction by using metrics including f1-score and ROUGE-2, similar to the Tianchi knowledge graph construction competition[1].
2. **Report Accuracy and Consistency**: We use metrics from natural language generation and image caption tasks to evaluate the report accuracy and consistency compared to the reference reports, including BLUE [22], METEOR [1], ROUGE [18], CIDEr [27], BERTScore [33].

[1] https://tianchi.aliyun.com/competition/entrance/532080/information

6 Results

As this paper is part of an early-stage Ph.D. symposium, we are currently in the process of running baseline models. Consequently, we do not yet have any results to report at the time of writing. However, we can refer to the results in reference [19], which indicate that knowledge enhancement and cross-modal approaches have a positive impact compared to image caption methods.

7 Conclusions/Lessons Learned

In conclusion, this paper has explored the challenges and limitations that current approaches are facing for radiology report generation, highlighting the critical issue of the lack of domain-specific knowledge in existing vision-language multimodal models. To address this research gap, we propose the integration of a knowledge enhancement module into existing vision-language multimodal models. The KEM module aims to extract and incorporate domain-specific knowledge from the visual embedding of radiograph features and medical knowledge graphs, enhancing the model's understanding of medical concepts and improving the quality of generated reports. By leveraging advanced AI techniques such as multimodal learning and knowledge fusion, the proposed approach seeks to overcome the limitations of current methods and pave the way for more effective and clinically useful radiology report generation systems. However, since this research is at an early stage, we are looking forward to discussing the research challenges we try to address in this paper and other challenges we may not be aware of, possible improvements to our proposed architecture and modules (Figures 1 and 2), as well as the implementation of an effective plan to evaluate our research.

References

1. 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)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
3. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
4. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79**, 102444 (2022)
5. Dalla Serra, F., Clackett, W., MacKinnon, H., Wang, C., Deligianni, F., Dalton, J., O'Neil, A.Q.: Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and

the 12th International Joint Conference on Natural Language Processing. pp. 615–624 (2022)

6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
7. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
8. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
9. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
11. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
12. Kale, K., Bhattacharyya, P., Jadhav, K.: Replace and report: NLP assisted radiology report generation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 10731–10742. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.683, https://aclanthology.org/2023.findings-acl.683
13. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
15. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
16. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021)
17. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018)
18. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Liu, C., Tian, Y., Song, Y.: A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199 (2023)

20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
21. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
22. 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)
23. Peng, C., Xia, F., Naseriparsa, M., Osborne, F.: Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review pp. 1–32 (2023)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Singh, S., Karimi, S., Ho-Shon, K., Hamey, L.: Show, tell and summarise: learning to generate and summarise radiology findings from medical images. Neural Computing and Applications **33**, 7441–7465 (2021)
26. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence **45**(1), 539–559 (2022)
27. 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)
28. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al.: Graph attention networks. stat **1050**(20), 10–48550 (2017)
29. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
30. Yasunaga, M., Ren, H., Bosselut, A., Liang, P., Leskovec, J.: Qa-gnn: Reasoning with language models and knowledge graphs for question answering. arXiv preprint arXiv:2104.06378 (2021)
31. You, J., Li, D., Okumura, M., Suzuki, K.: Jpg-jointly learn to align: Automated disease prediction and radiology report generation. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 5989–6001 (2022)
32. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
33. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr
34. Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C.D., Leskovec, J.: Greaselm: Graph reasoning enhanced language models. In: International conference on learning representations (2021)
35. Zhang, Y., Sui, E., Yeung-Levy, S.: Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. arXiv preprint arXiv:2401.08567 (2024)