ITAdaptor: Image-Tag Adapter Framework with Knowledge Enhancement for Radiology Report Generation

Shuaipeng Ding, Mengnan Fan, Mingyong Li(✉) and Cao Wang

School of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
limingyong@cqnu.edu.cn

Abstract. Automated radiology report generation holds significant research value as it has the potential to alleviate the heavy burden of report writing for radiologists. Previous studies have incorporated diagnostic information through multi-label classification to assist in report generation. However, these methods treate visual and diagnostic information equally, which overlooks the difference in the importance of both when generating different types of words. This can lead to errors in report generation. We propose the Image-Tag Adapter framework (ITAdaptor), which dynamically balances the contributions of visual and diagnostic information in the decoder, ensuring both are fully utilized during the report generation process. The model introduces two novel modules: Cross-Modal Knowledge Enhancement (CMKE) and Image-Tag Adapter (ITA). CMKE leverages pre-trained CLIP to retrieve similar reports from a database, assisting in the diagnosis of query images by providing relevant disease information. ITA adaptively fuses the visual information from the input images with the diagnostic information from the disease tags to generate more accurate reports. For training, we propose a strategy combining reinforcement learning and knowledge distillation, optimizing iteratively to extract knowledge into the ITAdaptor. Extensive comparative experiments on the IU-Xray and MIMIC-CXR benchmark datasets demonstrate the effectiveness of our proposed approach.

Keywords: Report Generation · Attention · Adapter · Knowledge Distillation

1 Introduction

Automated radiology report generation can significantly improve physician productivity and has therefore received increasing research attention. Mainstream approaches typically employ encoder-decoder architectures [17, 18, 26]. Early research used convolutional neural networks (CNNs) to extract visual features [24, 19]. With the advent of Transformer model [23], many studies have leveraged various attention mechanisms to enhance performance [9, 30]. In recent years, several methods, such as template retrieval structures [5, 13], memory-driven

networks [4, 3], and knowledge-aware modules [11, 15, 28], have shown promising results in report generation. Additionally, some studies have adopted multi-task learning, utilizing radiograph classification information to assist in report generation [20, 22].

Despite some progress, challenges remain for methods aimed at extracting radiological knowledge to assist in report generation. First, features extracted by the encoder from different modalities exist in distinct representation spaces, leading to inconsistent representations of image and text features with the same underlying semantics. Recent work [8] has somewhat alleviated this issue by distilling clinical information into the decoder. Secondly, in the decoder, it is unreasonable to treat visual and diagnostic information equally when generating different types of words by directly using disease classification results to assist in generation. For instance, diagnostic information plays a more crucial role when generating descriptions of abnormalities, such as pleural effusion and scoliosis.

Motivated by the limitations mentioned above, we propose ITAdaptor to enhance the utilization of radiographs and diagnostic knowledge, thereby improving automated report generation. Specifically, based on the encoder-decoder architecture, ITAdaptor incorporates a disease classification branch. During report generation, the diagnostic results from this branch are converted into disease tags to explicitly guide the generation process. To further improve diagnostic accuracy, we design Cross-Modal Knowledge Enhancement (CMKE), which leverages a pre-trained CLIP model to retrieve similar reports from the database, assisting in the diagnosis of query images. Additionally, we introduce the Image-Tag Adapter (ITA), which dynamically adjusts the weight of visual and diagnostic information to ensure that both advantages are fully utilized during report generation. Our main contributions are summarized as follows:

- We introduce a cross-modal knowledge enhancement that improves the representational capability of disease classification features by retrieving the most similar features, while also incorporating medical text knowledge into the report generation process, similar to how doctors consult relevant clinical records.
- We propose the Image-Tag Adapter, which utilizes visual and diagnostic information in the decoder to guide the report generation process, adaptively balancing the contributions of both based on the type of words being generated.
- To better align visual and textual features for generating radiology reports, we propose a three-stage training strategy that combines reinforcement learning and knowledge distillation, which utilizes iterative optimization to distill global knowledge into our ITAdaptor.

2 Method

In this section, the proposed model ITAdaptor is introduced. As shown in Fig. 1, the model includes two new modules: RKE and ITA. In addition, the disease classification branch serves as a disease tag generator to guide report generation.

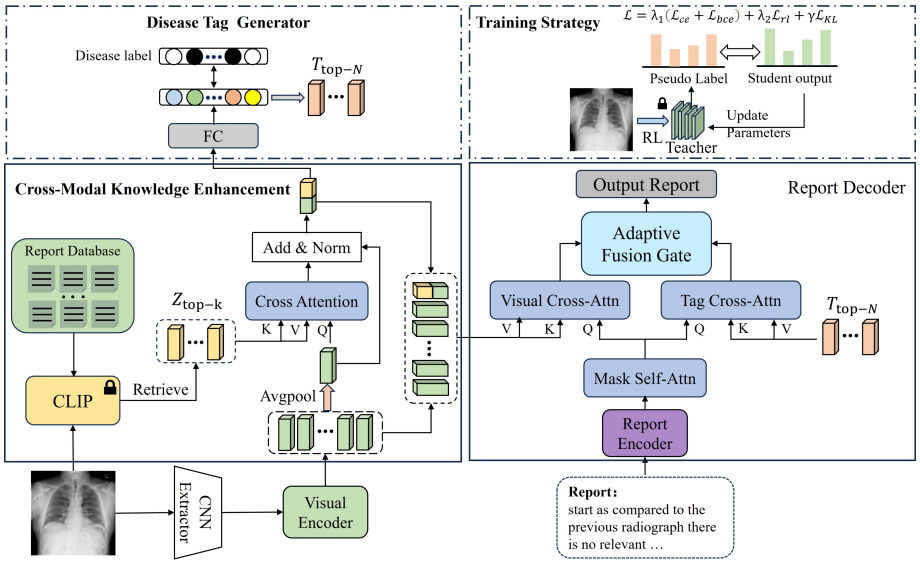

Fig. 1. ITAdaptor consists of three components: Cross-Modal Knowledge Enhancement (CMKE), Image-Tag Adapter (ITA) and Disease Tag Generator (i.e., Disease Classification Branch).

2.1 Cross-Modal Knowledge Enhancement

Diagnosing based solely on medical images may be suboptimal, as radiologists often have access to additional documentation for reference, such as patient information and diagnostic databases. Inspired by this, we leverage the report database from the training data to enhance the representational capacity of visual features and obtain more robust disease classification features.

We first utilize the CLIP model [6] pretrained on the MIMIC training set to perform cross-modal retrieval, aiming to retrieve the $top-k$ most relevant report embeddings Z_{top-k} for the input image. Given the input image I and the report library $\mathrm{R} = \{R_1, R_2, ...R_N\}$, where R_i represents the i-th radiology report and N is the number of reports, CLIP acts as an encoder that maps I and R_i to D-dimensional embeddings:

$$v = E_{image}\left(I\right), r_i = E_{text}\left(R_i\right) \tag{1}$$

Then we apply L2 normalization to standardize the embeddings and compute the cosine similarity between them to retrieve the most relevant report embeddings:

$$\mathrm{sim}\left(v, r_i\right) = \frac{v^T r_i}{\|v\| \, \|r_i\|} \tag{2}$$

$$\mathrm{Z}_{top-k} = \underset{r_i \in \mathrm{R}'}{\arg\max}^{top-k} \mathrm{sim}\left(v, r_i\right) \tag{3}$$

where R′= $\{r_1, r_2, ..., r_N\}$ is the set of report embeddings.

For the output features V of the visual encoder, we use average pooling to further aggregate features to obtain the feature V_g. We introduce multi-head cross attention ($MHCA$) to effectively integrate visual features with retrieved medical knowledge, which not only enhances cross-modal interaction but also yields more robust features for disease classification.

$$V_{CA} = MHCA\left(V_g, \mathrm{Z}_{top-k}\right) \tag{4}$$

then we follow [22] to further predict the disease tags T for the input image. Specifically, we feed V_{CA} into a multi-label classification network, which is pre-trained as a multi-label classification task on the downstream dataset to generate the distribution of all predefined disease tags. Finally, the most likely disease tags $D_{top-N} = \{d_1, d_2, ..., d_N\} \in \mathbb{R}^{N \times D}$ are used as the disease tags for the current input image.

2.2 Image-Tag Adapter

Decoder at each time step t, to generate each word y_t in the final report, our model first takes as input the embedding of the current input word $x_t = w_t + e_t$ (w_t : word embedding, and e_t : fixed position embedding) through multi-head self-attention (MHA) to obtain the current hidden state: $h_t = MHA(x_t, x_{1:t}) \in \mathbb{R}^D$. Then, visual cross attention and tag cross attention are introduced to capture salient visual information $v_t \in \mathbb{R}^D$ and salient diagnostic information $d_t \in \mathbb{R}^D$:

$$V_t = Visual\ Attention\left(h_t, [V; V_{CA}]\right) = MHCA\left(h_t, [V; V_{CA}]\right) \tag{5}$$

$$D_t = Tag\ Attention\left(h_t, D_{top-N}\right) = MHCA\left(h_t, D_{top-N}\right) \tag{6}$$

where[;] stands for concatenation operation.

When the decoder generates descriptions related to abnormalities and their severity, D_t is more important because it contains clear abnormal information. In contrast, when describing normal conditions as well as the location and shape of abnormalities, V_t is more critical because it encompasses the overall visual information. Therefore, we introduce an adaptive fusion gate to dynamically adjust the balance between the two parts:

$$\gamma_t = \sigma\left(F\left(D_t\right) - F\left(V_t\right)\right) \tag{7}$$

$$c_t = \gamma_t D_t + \left(1 - \gamma_t\right) V_t \tag{8}$$

Where σ is the sigmoid function, γ_t represents the importance of D_t compared to V_t, and F denotes the two fully connected layers used as scoring functions to evaluate the importance of diagnostic information and visual information. Finally the c_t is projected onto the vocabulary distribution through a fully connected layer and a softmax function.

2.3 Training Strategy

The entire training process consists of three stages:

First, the ITAdaptor model is pretrained using word level cross entropy loss $\mathcal{L}_{ce}$ and binary cross entropy loss $\mathcal{L}_{bce}$ for multi-label classification:

$$\mathcal{L}_{ce} = -\frac{1}{N^r}\sum_{i=1}^{N^r} w_i \cdot \log\left(p_i\right) \tag{9}$$

$$\mathcal{L}_1 = \mathcal{L}_{ce} + \mathcal{L}_{bce} \tag{10}$$

where $\{p_i\}$is the predicted sequence of word markers and $\{w_i\}$ is the corresponding ground truth report.

Second, inspired by CMM+RL [21], after the model has been trained with several epochs, the sequence generation is fine-tuned by Reinforcement Learning, and reinforcement learning loss is defined as: $\mathcal{L}_{rl} = \nabla_\theta L_\theta = -(r(\omega) - b)\nabla_\theta \log\left(p_\theta(\omega)\right)$, where r(.) refers to the reward function, and b refers to the reward value, which is obtained from BLEU-4 metric. The training objective of the current stage is the combination of $\mathcal{L}_1$ and $\mathcal{L}_{rl}$, with $\mathcal{L}_1$ and $\mathcal{L}_{rl}$ scaled by factors λ_1 and λ_2:

$$\mathcal{L}_2 = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_{rl} \tag{11}$$

Third, we use the models trained in the first and second stages as the student and teacher models, respectively, and apply knowledge distillation (KD) to minimize the difference between their probability distributions for word index c and image I, expressed as: $\mathcal{L}_{KL} = \frac{1}{N}\sum_{c=1}^{N} KL\left[p_t(c,I)\|p_s(c,I)\right]$, N is the dimension of the word space. The final stage of training is the combination of $\mathcal{L}_2$ and $\mathcal{L}_{KL}$,with $\mathcal{L}_{KL}$ scaled by factor γ:

$$\mathcal{L}_3 = \mathcal{L}_2 + \gamma\mathcal{L}_{KL} \tag{12}$$

Specifically, when the student model performs better than the teacher network in a new epoch, we transfer the weights of the student to the teacher and iteratively conduct the training of the third stage.

3 Experiments

Datasets and Evaluation Metrics. We evaluate our model on two widely-used datasets for report generation: IU-Xray [5] and MIMIC-CXR [10]. IU-Xray dataset, developed by Indiana University, is a dataset containing 7,470 X-ray images and 3,955 corresponding reports. We follow the established training-validation-testing splits of previous research [14, 2] with a distribution ratio of 7: 1 : 2. MIMIC-CXR dataset, released by BethIsrael Deaconess Medical Center, is a comprehensive chest X-ray dataset containing 473,057 radiographs and 206,563 corresponding reports. Following previous works [3, 4], we utilize the official split, where the training set consists of 368,960 images, the validation set contains 2,991 images, and the test set contains 5,159 images.

For both datasets, we used categorical labels from [22] for the classification task. We assessed the quality of the generated reports using various evaluation metrics. These include BLEU [5], METEOR [1] and ROUGE-L [12]. Higher scores are indicative of superior model performance.

Table 1. Performance comparisons of the proposed ITAdaptor with existing methods on NLG metrics were conducted using the test sets of the MIMIC-CXR and IU-Xray datasets. optimal and suboptimal performance is highlighted.

Dataset	Methods	Venue	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L
IU-Xray	R2Gen	EMNLP'20	0.470	0.304	0.219	0.165	-	0.371
	R2GenCMN	ACL'21	0.475	0.309	0.222	0.170	0.375	0.191
	GSKET	MedIA'22	0.496	0.327	0.238	0.178	-	
	M2KT	MedIA'23	0.497	0.319	0.230	0.174	-	0.399
	GMoD	MICCAI'24	0.530	0.363	0.267	0.203	0.217	0.418
	EKAGen	CVPR'24	0.526	0.361	0.267	0.203	0.214	0.404
	RAMT	TMM'24	0.482	0.310	0.221	0.165	0.195	0.377
	ITAdaptor	Ours	0.536	0.377	0.274	0.206	0.220	0.420
MIMIC-CXR	R2Gen	EMNLP'20	0.353	0.218	0.145	0.103	0.142	0.270
	R2GenCMN	ACL'21	0.353	0.218	0.148	0.106	0.142	0.278
	GSKET	MedIA'22	0.363	0.228	0.156	0.115	-	0.284
	M2KT	MedIA'23	0.386	0.237	0.157	0.111	0.137	0.274
	GMoD	MICCAI'24	0.398	0.251	0.172	0.124	0.166	0.286
	EKAGen	CVPR'24	0.419	0.258	0.170	0.119	0.157	0.287
	RAMT	TMM'24	0.362	0.229	0.157	0.113	0.153	0.284
	ITAdaptor	Ours	0.411	0.260	0.187	0.141	0.152	0.314

Implementation Details. Our baseline model includes a pre-trained ResNet 101 [7], a 3-layer Transformer encoder, a 3-layer Transformer decoder, and an additional disease classification branch. the $top-N$ of ITA is 5. The loss scaling factors λ_1, λ_2, and γ are set to 0.01, 0.99, and 0.01, respectively. We use the AdamW [16] optimizer, with a learning rate of 2 $\times$ 10-5 for the Visual Extractor and 1 $\times$ 10-4 for the language generation model. The training batch sizes for MIMIC-CXR and IU-Xray are set to 32 and 16, respectively. All experiments are conducted on an RTX 4090 GPU.

4 Analysis

Performance Comparison. We compared our method with several SOTA methods using the MIMIC-CXR and IU-Xray datasets. The selected comparison methods include Knowledge-Based methods (GSKET [28], M2KT [27], EKAGen [2]), the method using graph structures (GMoD [25], RAMT [29]), and Memory Driven methods (R2Gen [4], R2GenCMN [3]). As shown in Table 1, our method achieved state-of-the-art performance on the IU-Xray dataset and competitive results on the MIMIC-CXR dataset, with scores 1.5% and 1.7% higher than the suboptimal model on the BL-3 and BL-4 metrics, respectively. This suggests that our model captures contextual information and word relationships better. Furthermore, our model outperformed the suboptimal model by 2.4% on the ROUGE-L metric, demonstrating its superior ability to generate key phrases.

Table 2. Ablation studies on the proposed Cross-Modal Knowledge Enhancement (CMKE), Image-Tag Adapter (ITA) and Training Strategy (TS). In TS, RL stands for reinforcement learning, KD stands for knowledge distillation.

Dataset	CMKE	ITA	TS		Metric					
			RL	KD	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L
MIMIC-CXR	✗	✗	✗	✗	0.348	0.213	0.144	0.102	0.133	0.271
	✓	✗	✗	✗	0.369	0.229	0.156	0.112	0.143	0.284
	✗	✓	✗	✗	0.382	0.241	0.165	0.118	0.148	0.284
	✓	✓	✗	✗	0.395	0.253	0.170	0.125	0.151	0.288
	✓	✓	✓	✗	0.387	0.253	0.181	0.136	0.141	0.294
	✓	✓	✓	✓	**0.411**	**0.260**	**0.187**	**0.141**	**0.152**	**0.314**

Ablation Studies. As shown in Table 2, first, the Base+CMKE showed improvements on all metrics, indicating that absorbing medical expertise from similar cases is beneficial. Second, the Base+ITA achieved more significant enhancements, with increases of 3.4% and 1.6% in BL-1 and BL-4, respectively. This demonstrates the effectiveness of adaptively determining which information to rely on when generating words to produce disease-oriented visual features. Notably, CMKE can obtain more robust features for classification, resulting in more accurate disease tags for ITA. When combining CMKE and ITA, performance further improves.

For RL and KD, the model+RL showed a decrease in BL-1 and METEOR, while improving by 1.1% in both BL-3 and BL-4. When RL and KD are combined simultaneously, the model achieves the best overall performance, indicating the effectiveness of the three-stage training strategy.

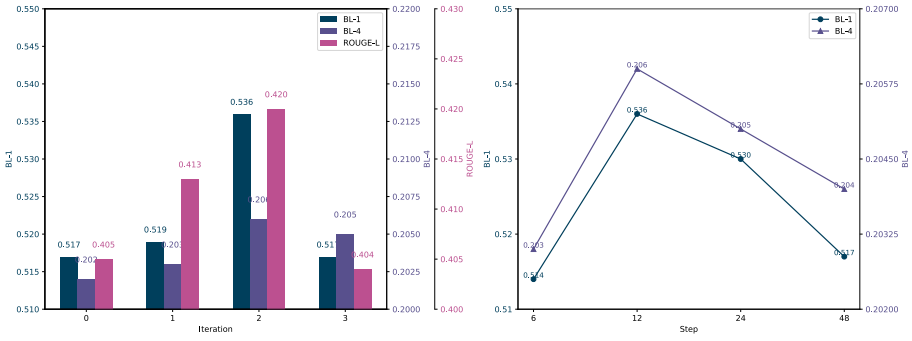

Fig. 2. (a) The impact of the number of iterations in the third stage of training on IU-Xray. (b) Effect of varying top-k on IU-Xray.

Additionally, we analyzed the impact of the number of iterations in the third stage of training on report generation performance on the IU-Xray dataset. Fig. 2 (a) shows that the model performed best after two optimization iterations. We also examined the effect of different $top-k$ values on report generation performance. $top-k$ is a key hyperparameter of the CMKE module that determines the number of relevant reports retrieved during the current report generation process. Excessive relevant reports may introduce noise. Fig. 2 (b) shows that the model achieved peak performance at $top-k=12$.

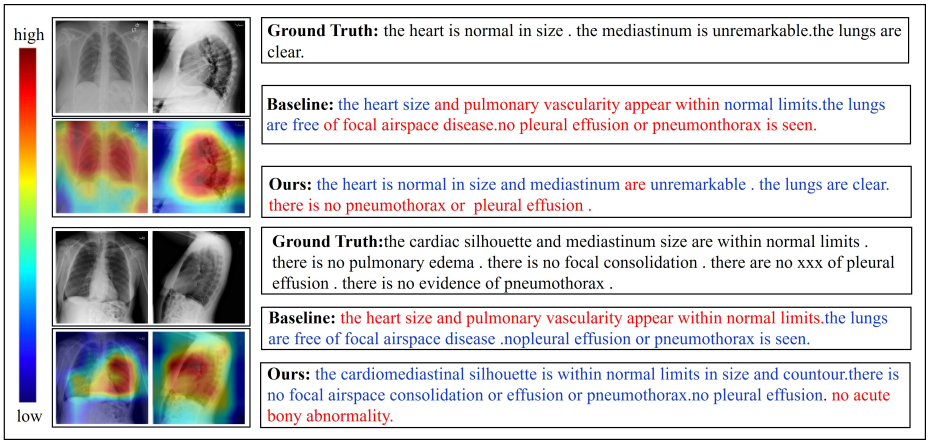

Fig. 3. Visualization: the network is highly concerned about this red area, the blue area that is not concerned. Report: correct descriptions are highlighted in blue, while incorrect descriptions are shown in red.

Quantitative Analysis. We draw attention maps to explore the regions of the medical images that the generated reports focus on. Fig. 3 shows that our model accurately identifies the target areas and produces reports that are closer to the ground truth. This suggests that the model can reduce the cross-modal gap and generate reports that are more consistent with the images.

5 Conclusion

This paper proposes ITAdaptor, a novel architecture dedicated to enhancing information utilization. Our method can dynamically adjust the contributions of visual information and diagnostic information in report generation based on the type of generated words, fully leveraging the advantages of both, and combining the training strategy to iteratively optimize report generation. Significant improvements across MIMIC-CXR and IU-Xray illustrate the effectiveness and

generation of our proposed method. Although we have made progress in radiology report generation, there are still some limitations. In future work, we will focus on validating the generalizability and effectiveness of our proposed model through a real-world pilot study across local hospitals.

Acknowledgements. This work was supported by the Science and Technology Research Project of Chongqing Education Commission (KJQN202300514), the Key Project of Innovation and Development Joint Fund of Chongqing Natural Science Foundation (CSTB2023NSCQ-LZX0003).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

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. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14204 (2024)
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
5. 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)
6. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health. pp. 209–219. PMLR (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19809–19818 (2023)
9. Ji, J., Luo, Y., Sun, X., Chen, F., Luo, G., Wu, Y., Gao, Y., Ji, R.: Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1655–1663 (2021)
10. 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)

11. Li, M., Liu, R., Wang, F., Chang, X., Liang, X.: Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web **26**(1), 253–270 (2023)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
13. Liu, F., Ge, S., Zou, Y., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. arXiv preprint arXiv:2206.14579 (2022)
14. Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. arXiv preprint arXiv:2106.06965 (2021)
15. Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al.: Auto-encoding knowledge graph for unsupervised medical report generation. Advances in Neural Information Processing Systems **34**, 16266–16279 (2021)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 375–383 (2017)
18. Ma, S., Han, Y.: Describing images by feeding lstm with structural words. In: 2016 IEEE international conference on multimedia and expo (ICME). pp. 1–6. IEEE (2016)
19. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
20. Pan, R., Ran, R., Hu, W., Zhang, W., Qin, Q., Cui, S.: S3-net: A self-supervised dual-stream network for radiology report generation. IEEE Journal of Biomedical and Health Informatics **28**(3), 1448–1459 (2023)
21. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 448–458 (2022)
22. Shang, C., Cui, S., Li, T., Wang, X., Li, Y., Jiang, J.: Matnet: Exploiting multimodal features for radiology report generation. IEEE Signal Processing Letters **29**, 2692–2696 (2022)
23. 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)
24. 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)
25. Xiang, Z., Cui, S., Shang, C., Jiang, J., Zhang, L.: Gmod: Graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 295–305. Springer (2024)
26. Xu, K., Wang, H., Tang, P.: Image captioning with deep lstm based on sequential residual. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 361–366. IEEE (2017)
27. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023)

28. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical image analysis **80**, 102510 (2022)
29. Zhang, K., Jiang, H., Zhang, J., Huang, Q., Fan, J., Yu, J., Han, W.: Semi-supervised medical report generation via graph-guided hybrid feature consistency. IEEE Transactions on Multimedia **26**, 904–915 (2023)
30. Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15465–15474 (2021)