GMoD: Graph-driven Momentum Distillation Framework with Active Perception of Disease Severity for Radiology Report Generation

ZhiPeng Xiang[1], ShaoGuo Cui[2⋆], CaoZhi Shang[3], Jingfeng Jiang[4], Liqiang Zhang[5]

[1,2]Chongqing Normal University, China
zhipengxiang@yeah.net; csg@cqnu.edu.cn
[3]Huazhong University of Science and Technology, China
[4]Michigan Technological University, USA
[5]The First Afliated Hospital of Chongqing Medical University, Chongqing, China

Abstract. Automatic radiology report generation is a challenging task that seeks to produce comprehensive and semantically consistent detailed descriptions from radiography (e.g., X-ray), alleviating the heavy workload of radiologists. Previous work explored the introduction of diagnostic information through multi-label classification. However, such methods can only provide a binary positive or negative classification result, leading to the omission of critical information regarding disease severity. We propose a Graph-driven Momentum Distillation (GMoD) approach to guide the model in actively perceiving the apparent disease severity implicitly conveyed in each radiograph. The proposed GMoD introduces two novel modules: Graph-based Topic Classifier (GTC) and Momentum Topic-Signal Distiller (MTD). Specifically, GTC combines symptoms and lung diseases to build topic maps and focuses on potential connections between them. MTD constrains the GTC to focus on the confidence of each disease being negative or positive by constructing pseudo labels, and then uses the multi-label classification results to assist the model in perceiving joint features to generate a more accurate report. Extensive experiments and analyses on IU-Xray and MIMIC-CXR benchmark datasets demonstrate that our GMoD outperforms state-of-the-art method. Our code is available at https://github.com/xzp9999/GMoD-mian.

Keywords: Report Generation · Attention · Knowledge Distillation.

1 Introduction

Automated generation of radiology reports can improve physician productivity and has received increasing research attention. Mainstream approaches use encoder-decoder architectures [1,22,21,33,30,7,20]. Early research used convolutional neural networks (CNN) to extract visual features [1,22,30]. Also, Long

[⋆] Corresponding Author.

Short-Term Memory Networks (LSTM) [21,33,30] or recurrent neural networks (RNN) [1,22] were also used to generate reports or extract features based on regions [39]. With the advent of the Transformer [27], many studies have used various attention mechanisms to improve the performance [7,20,39]. Recently, some methods have started exploring the extraction of radiological knowledge to assist in report generation [31,36,35,12,17,40,6], with [36] proposing a knowledge-enhanced attention mechanism to improve generation quality. with some studies employing multitask learning and utilizing radiograph classification information to aid in report generation [25,23].

Despite some progress, challenges remain for methods aiming to extract radiological knowledge to assist in report generation. First, features extracted by the encoder from different modalities exist in different representation spaces, leading to inconsistent representations of image and text features with the same semantics. Recent work [6] distilling clinical information into decoders can somewhat alleviate this problem. Secondly, directly using disease or symptom classification results to assist in generation (e.g., emphysema-positive) lacks consideration for the degree of negativity or positivity (e.g., lung field transparency increases, probably emphysema). Lastly, the interconnections between diseases and symptoms, as well as between different diseases, are disregarded.

Motivated by the shortcomings mentioned above, we propose GMoD to enhance the utilization of radiographs and diagnostic knowledge for improved automated report generation. Our framework mainly consists of two core modules, GTC and MTD. First, we construct a similarity graph between disease and symptom features through pre-training. Then, with the help of the graph attention mechanism, we guide the model to focus on the potential relationships between symptoms and pathological features, thereby enhancing the model's understanding of the images and improving the quality of report generation. MTD is designed to assess the degree of disease negativity or positivity by creating non-one-hot pseudo-labels to guide the model in learning these nuances, instead of rigidly categorizing them into two categories. Our contributions can be summarized as follows: 1) To make full use of medical information we propose the GMoD framework with two novel modules: Graph-based Topic Classifier (GTC) and Momentum Topic-Signal Distiller (MTD); 2) GTC guides the model to emphasize potential relationships between symptoms and pathological features during classification tasks for report generation. While MTD constructs pseudo-labels on this basis and adds distillation loss constraints to the classifier to focus on the confidence of disease negativity or positivity; 3) Extensive experiments and research conducted on the MIMIC-CXR and IU-Xray datasets demonstrate that our GMoD has achieved state-of-the-art performance.

2 Methodology

As shown in Fig. 1, our proposed GMoD framework contains two novel modules: GTC and MTD. Notably, MTD shares GTC's structure. MTD's weight initialization is copied from GTC without gradient, updated by momentum in

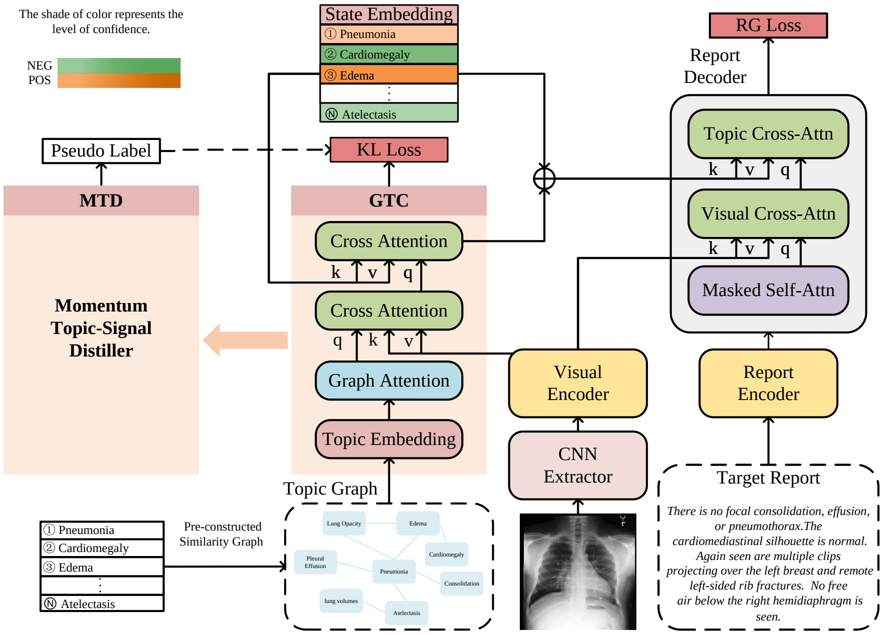


Fig. 1. The overall architecture of GMoD, where the MTD module is identical to that of the GTC, and the initialization parameters are passed through from the GTC.

subsequent calculations. The pseudo-label for GTC loss is constructed using the output of MTD and ground truth.

2.1 Graph-based Topic Classifier

We adopted a graph attention mechanism to capture the intrinsic correlation between diseases and symptoms. Specifically, we extracted 100 symptom topic headings and 14 lung diseases, jointly constructed a topic graph, and then vectorized it using topic embeddings. We chose a two-stage training approach, as described below.

Creating a pre-constructed similarity graph is the objective of the first stage. The vectors $V = \{v_1, v_2, \ldots, v_n\}$ from the topic embedding were first randomly initialized, where $V \in R^{n \times d}$. Subsequently, we used a classifier without graph attention to update the topic embedding, aiming to learn an embedding effectively representing disease features. The trained embedding was used to construct a similarity graph and a similarity matrix was evaluated using the formulation below:

$$Sim_{(i,j)} = \frac{V_i \times V_j}{||V_i|| \times ||V_j||}, \quad i, \ j \in \{1, 2, ..., n\} \tag{1}$$

where n denotes the number of features in the topic embedding. Then, the top K most similar features (from the similar matrix; see Eqn. 1) were selected to form a graph.

A Graph Classifier provides disease and symptom diagnostic information to the decoder for report generation tasks as the goal of the second stage. Vectors V_G were obtained from pre-trained topic embedding. The process of learning the relationship between diseases and symptoms through the graph attention mechanism (GAT) [29] can be written as $F_g = GAT(V_G)$. Cross-Attention Mechanism (CAM) can be defined below:

$$CAM(X, Y) = \alpha Y W^V, \alpha = softmax(XW^Q(YW^K)^T). \tag{2}$$

where W^Q, W^K and W^V are learnable parameters. The divisor $\sqrt{d}$ is omitted in the above formula for simplicity. After receiving the graph features $F_g \in \mathbb{R}^{n\times d}$, we employed the cross-attention mechanism to capture the radiographic regions of interest for each node in the topic graph $F'_\nu = CAM(F_\nu, F_g)$, where $F_v \in \mathbb{R}^{N\times d}$ represents the radiographic features extracted by the image encoder and $F'_\nu \in \mathbb{R}^{n\times d}$ is the fused feature of symptoms graph and radiographs. The process of utilizing cross-attention to focus on the features F_s in the State Embedding is represented as: $F'_s = CAM(F'_v, F_s) = \alpha_c F_s W^V$. Here, F'_s represents the diagnostic features that obtained the classification results, while α_c serves as the attention score, functioning as the probability distribution used for calculating the MTD loss. Then, we use a method similar to Shang et al.[25] to further obtain state-aware disease features.

$$\nu' = \begin{cases} 1, & \alpha_c > \tau \\ 0, & \alpha_c < \tau \end{cases}, F_c = \begin{cases} F'_s + yF_s, & \text{if training phase} \\ F'_s + y'F_s, & \text{otherwise} \end{cases} \tag{3}$$

where $y \in \mathbb{R}^{n\times 2}$ is one-hot ground-truth label, and $y' \in R^{n\times 2}$ is the prediction result, and τ is a hyperparameter serving as a threshold. In Eqn. 3, $S \in \mathbb{R}^{2\times d}$ represents state embedding, which is randomly initialized and updated through training and $F_c \in \mathbb{R}^{n\times d}$ represents the state-aware disease features utilized to assist the decoder during the (radiology) report generation.

2.2 Momentum Topic-Signal Distiller

We observed that one-hot labels solely indicate the presence or absence of a disease or condition, overlooking the crucial information regarding the degree of negativity or positivity associated with the disease. In general, the more severe the disease depicted in the radiograph, the higher the confidence observed through the MTD module. Therefore, we employed momentum to generate pseudo-labels and leveraged self-distillation to help the model better understand the severity of each disease and symptom. Throughout the training process, MTD had no gradient. MTD parameters were updated as follows:

$$P_{MTD} \leftarrow mP_{MTD} + (1-m)P_{GTC} \tag{4}$$

where $m \in [0, 1)$ is a momentum coefficient, P_{GTC} and P_{MTD} are parameters of GTC and MTD modules, respectively. After updating P_{MTD}, we sent F_g and F_v into MTD for forward propagation again and constructed the pseudo labels using obtained α'_c and the classification ground truth label L^{GT}, as follows:

$$\alpha'_c, F'_c = MTD(F_v, F_g) \tag{5}$$

Where α'_c and F'_c are obtained in a manner equivalent to α_c and F_c in GTC. Then, pseudo labels were constructed using the obtained α'_c and the categorical true label L^{GT}, and the distiller loss was calculated:

$$P^T_\theta = \mu L^{GT} + (1 - \mu)\text{softmax}(\alpha'_c) \tag{6}$$

where $\mu \in$ [0, 1) is a distillation coefficient, P^T_θ represents the targeted probability distribution obtained by weighting the true labels and the momentum-predicted labels. We used Kullback–Leibler divergence to minimize the difference between P^T_θ and the probability distribution α_c, expressed as: $\mathcal{L}_{DC} = KL(P^T_\theta || \alpha_c)$. As the classifier optimization objective. The distillation-base classifier loss is denoted as $\mathcal{L}_{DC}$ in Eqn. 8. The distillation coefficient μ followed a classical decay strategy:

$$\mu = \mu - (1 - cos(\pi \cdot \frac{e_{cur}}{e_{tol}}) \cdot (\mu - \mu')) \tag{7}$$

where μ' is the minimum value to which the distillation coefficient should decay, e_{cur} represents the current epoch value, and e_{tol} denotes the total training epochs. The overall optimization objective consists of the report generation loss $\mathcal{L}_{RG}$ and distillation-based classifier loss $\mathcal{L}_{DC}$. λ_{DC} and λ_{RG} are loss weighting hyper-parameters.

$$\mathcal{L} = \lambda_{DC} \cdot \mathcal{L}_{DC} + \lambda_{RG} \cdot \mathcal{L}_{RG} \tag{8}$$

3 Experiments

Dataset And Evaluation Metrics. Through extensive experiments on the widely used datasets MIMIC and IU-Xray, we thoroughly evaluate the proposed GMoD framework. **MIMIC-CXR[8]** provided by Beth Israel Deaconess Medical Center, is a recently released large-scale data set. The dataset includes 377,110 radiographs and 227,835 diagnostic reports. For a fair comparison, we used the same data splits and Vocabulary Building as the benchmark method[3], resulting in 368,960 in the training set, 2,991 in the validation set, and 5,159 in the test set. **IU-Xray[5]** supplied by Indiana University contains 7,470 X-ray images and 3,955 corresponding reports. Following previous work [3], we excluded a portion of data without "Findings" and split 70%-10%-20% training-validation-testing sets, following the most widely used practices.

For both datasets, we used categorical labels from [25] for the classification task, with the occurrence of the topic words being noted as positive and vice versa as negative. We assessed the quality of the generated reports using various evaluation metrics. These include BLEU [5], METEOR [2], ROUGE-L [15], and CIDEr [28]. Higher scores are indicative of superior model performance.

Table 1. Performance comparisons of the proposed GMoD with existing methods on NLG metrics were conducted using the test sets of the MIMIC-CXR and IU-Xray datasets. The best values are highlighted in bold.

Datasets	Methods	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
IU-Xray	R2Gen [3]	2020	0.470	0.304	0.219	0.165	0.187	0.371	-
	PPKED [17]	2021	0.483	0.315	0.224	0.168	0.190	0.376	0.351
	Contrastive [18]	2021	0.492	0.314	0.222	0.169	0.193	0.381	-
	AlignTransformer [37]	2021	0.484	0.313	0.225	0.173	0.204	0.379	-
	CMCL [16]	2022	0.473	0.305	0.217	0.162	0.186	0.378	-
	URA [14]	2023	0.530	**0.365**	0.263	0.200	0.218	0.405	**0.510**
	KiUT [6]	2023	0.525	0.360	0.251	0.185	**0.242**	0.409	-
	GMoD	Ours	**0.530**	0.363	**0.267**	**0.203**	0.217	**0.418**	0.437
MIMIC -CXR	TopDown [1]	2018	0.317	0.195	0.130	0.092	0.128	0.267	-
	R2Gen [3]	2020	0.353	0.218	0.145	0.103	0.142	0.270	-
	PPKED [17]	2021	0.360	0.224	0.149	0.106	0.149	0.284	0.237
	Contrastive [18]	2021	0.350	0.219	0.152	0.109	0.151	0.283	-
	AlignTransformer [37]	2021	0.378	0.235	0.156	0.112	0.158	0.283	-
	CMCL [16]	2022	0.344	0.217	0.140	0.097	0.133	0.281	-
	URA [14]	2023	0.363	0.229	0.158	0.107	0.157	0.286	0.246
	KiUT [6]	2023	0.393	0.243	0.159	0.113	0.160	0.285	-
	GMoD	Ours	**0.398**	**0.251**	**0.172**	**0.124**	**0.166**	**0.286**	**0.377**

Implementation Details. Our baseline model utilized the pre-trained DenseNet-121 to extract image features. A 6-layer image-text encoder and a 12-layer report decoder for report generation complemented it. The image encoder used a self-attention mechanism, while the text encoder used a masked self-attention mechanism. We used an Adam[19] optimizer with a weight decay of 5e-5, and set the learning rate to 1e-4. The top k highest similarities were chosen as interconnected nodes to obtain the pre-constructed similarity graph, with k set to 5. Distillation coefficient μ and μ' were set to 0.995 and 0.95 for the MIMIC data, respectively. The same parameters for the IU-Xray were 0.995 and 0.96, respectively. The momentum coefficient m was set to 0.995. Larger generation loss weights yield better results; therefore, λ_{DC} and λ_{RG} were set to 2 and 1, respectively. To be consistent with existing methods, we simultaneously used frontal and lateral radiology images as input on the IU-Xray dataset and each image separately as input on MIMIC.

4 Analysis

Performance Comparison. To demonstrate our architecture's effectiveness, we compared several selected state-of-the-art methods using the MIMIC and IU-Xray datasets. The selected comparison methods included a baseline method[1,3], a method with improved attention mechanisms [18,37], and a method using graph structures [6]. As shown in Table 1, our GMoD architecture achieved competitive results in both datasets. Furthermore, our results indicated that graph-driven knowledge distillation could more effectively integrate diagnostic information for improved (radiology) report generation.

Table 2. Ablation studies on the proposed Graph-based Topic Classifier (GTC) and Momentum Topic-Signal Distiller (MTD). In GTC, SC stands for a standalone classifier without a topic graph, while GA represents building a topic graph on this basis and adding graph attention.

dataset	GTC		MTD	Metric						
	SC	GA		Bleu1	Bleu2	Bleu3	Bleu4	Meteor	Rouge_L	CIDEr
MIMIC-CXR				0.356	0.222	0.151	0.108	0.140	0.280	0.253
	✓			0.378	0.228	0.148	0.099	0.149	0.270	0.224
	✓	✓		0.398	0.239	0.155	0.104	0.156	0.269	0.304
	✓		✓	0.392	0.246	0.169	0.122	0.163	**0.287**	0.359
	✓	✓	✓	**0.398**	**0.251**	**0.172**	**0.124**	**0.166**	0.286	**0.377**

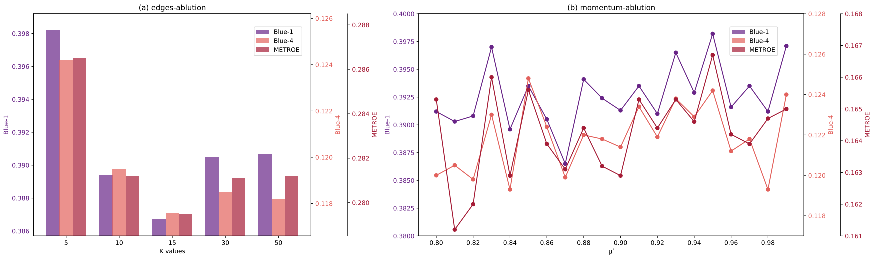


Fig. 2. (a) The effect of different values of K when choosing the top K most similar Topic words as nodes for interconnection. (b) Experimental results corresponding to the minimum value μ' of different momentum coefficient attenuation

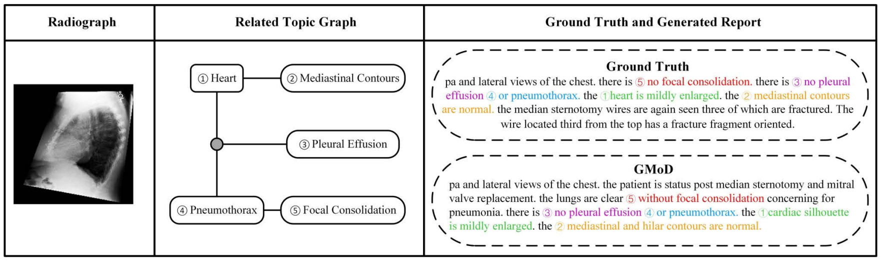
Ablation Studies. To thoroughly explore the contributions of our proposed GTC and MTD modules, our major results are shown in Fig. 2. In the ablation study, GTC and MTD were added separately to the baseline model.

For GTC, we started with the baseline model and first added a standalone classifier without a topic graph (SC). Then, we introduced a pre-constructed topic graph and added graph attention (GA) based on SC to train the classifier. Results in Table 2 clearly indicate that adding both SC and GA improved the performance, underscoring the effectiveness of guiding the model to learn and leverage this implicit relationship for enhancing report generation.

When constructing a similarity graph and selecting the top K most similar topic words as nodes, a larger K value requires the model to focus on more nodes with low correlation. As a result, introducing a large K introduces noise and may cause the model to lose attention to important nodes. Fig. 2 (a) illustrates the impact of different K values on report generation performance.

For MTD, From Table 2, for the MIMIC dataset, we can observe that MTD significantly improves the quality of generated reports. The report is generated by refining the diagnostic information through MTD, which aligns more with the radiologist's decision-making process.

Through observations, we found that a larger distillation coefficient is more beneficial. Therefore, we set the initial value of μ to 0.995, and the impact of



Ground Truth

pa and lateral views of the chest. there is ⑤ no focal consolidation. there is ③ no pleural effusion ④ or pneumothorax. the ① heart is mildly enlarged. the ② mediastinal contours are normal. the median sternotomy wires are again seen three of which are fractured. The wire located third from the top has a fracture fragment oriented.

GMoD

pa and lateral views of the chest. the patient is status post median sternotomy and mitral valve replacement. the lungs are clear ⑤ without focal consolidation concerning for pneumonia. there is ③ no pleural effusion ④ or pneumothorax. the ① cardiac silhouette is mildly enlarged. the ② mediastinal and hilar contours are normal.

Fig. 3. A case study of our model utilizing a topic graph to generate radiology reports. The related topic graph shows the five nodes related to this example, where the grey circle refers to the global node. Different colors in the report indicate the radiology terms associated with the different nodes.

the minimum distillation coefficient μ' on the experimental results is shown in Fig. 2 (b).

Case Studies. Case studies illustrated in Fig. 3. It can be intuitively seen that our method precisely and correctly generates reports consistent with the ground truth. As Fig. 3 shows, our method correctly generates five radiological terms consistent with the ground truth, we attribute this improvement to the graph construction and Graph-based Topic Classifier, which accurately establishes connections between radiographs and radiological terms. At the same time, it accurately assesses their normal and abnormal conditions. Notably, our method attends to and reports some subtle observations, such as "cardiac silhouette is mildly enlarged", which can be attributed to the knowledge extracted by the distiller.

5 Conclusions

This paper proposes GMoD, a novel architecture dedicated to enhancing information utilization. For the first time, we attempted to align high-level visual information with disease severity via explicit constraints when generating radiology reports. Our model utilized the potential relationship between topic words and radiographs for encoding and established the mapping logic between disease severity and radiographs through the momentum distiller constraint model, thereby perceiving disease severity proactively. Extensive experiments on MIMIC and IU-Xray show that our method correctly established the connection between radiographs, diagnostic information, and reports, proving the effectiveness of auxiliary report generation focusing on disease severity.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (62306054), the Natural Science Foundation Project of Chongqing Science and Technology Bureau(CSTB2022NSCQ-MSX1206), Chongqing Postgraduate Scientific Research Innovation Project(CYS23406), the Key Science and Technology Research Program of Chongqing Municipal Education Commission (KJZD-K202200510).

The Technology Foresight and System Innovation Project of Chongqing Municipal Science and Technology Bureau(CSTB2022TFII-OFX0042)

Disclosure of Interests. we declare no competing interests.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: IEEvaluation@ACL (2005)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
4. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: CVPR, (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 (2016)
6. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: CVPR (2023)
7. 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: AAAI (2021)
8. 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)
9. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. arXiv preprint arXiv:2205.12005 (2022)
10. 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 (2022)
11. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems (2021)
12. Li, M., Cai, W., Verspoor, K., Pan, S., Liang, X., Chang, X.: Cross-modal clinical graph transformer for ophthalmic report generation. In: CVPR (2022)
13. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: CVPR, (2023)
14. Li, Y., Yang, B., Cheng, X., Zhu, Z., Li, H., Zou, Y.: Unify, align and refine: Multi-level semantic alignment for radiology report generation. In: CVPR, (2023)
15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out (2004)
16. Liu, F., Ge, S., Zou, Y., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. arXiv preprint arXiv:2206.14579 (2022)

17. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: CVPR (2021)
18. 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)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR (2017)
21. Ma, S., Han, Y.: Describing images by feeding lstm with structural words. In: ICME (2016)
22. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks. arXiv preprint arXiv:1412.6632 (2014)
23. 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 (2023)
24. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: ACL (2022)
25. Shang, C., Cui, S., Li, T., Wang, X., Li, Y., Jiang, J.: Matnet: Exploiting multimodal features for radiology report generation. IEEE Signal Processing Letters (2022)
26. Song, Z., Zhou, X.: Exploring explicit and implicit visual relationships for image captioning. In: ICME (2021)
27. 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 (2017)
28. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015)
29. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
31. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: ECCV (2022)
32. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: MICCAI (2022)
33. Xu, K., Wang, H., Tang, P.: Image captioning with deep lstm based on sequential residual. In: ICME (2017)
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning (2015)
35. 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 (2023)
36. 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 (2022)
37. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: MICCAI (2021)

38. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
39. 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: CVPR (2021)
40. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: AAAI (2020)