Utilizing Longitudinal Chest X-Rays and Reports to Pre-fill Radiology Reports

Qingqing Zhu[1], Tejas Sudharshan Mathai[2], Pritam Mukherjee[2], Yifan Peng[3], Ronald M. Summers[2], and Zhiyong Lu[1](✉)

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
zhiyong.lu@nih.gov

[2] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Department of Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD, USA

[3] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA

Abstract. Despite the reduction in turn-around times in radiology reporting with the use of speech recognition software, persistent communication errors can significantly impact the interpretation of radiology reports. Pre-filling a radiology report holds promise in mitigating reporting errors, and despite multiple efforts in literature to generate comprehensive medical reports, there lacks approaches that exploit the longitudinal nature of patient visit records in the MIMIC-CXR dataset. To address this gap, we propose to use longitudinal multi-modal data, i.e., previous patient visit CXR, current visit CXR, and the previous visit report, to pre-fill the "findings" section of the patient's current visit. We first gathered the longitudinal visit information for 26,625 patients from the MIMIC-CXR dataset, and created a new dataset called *Longitudinal-MIMIC*. With this new dataset, a transformer-based model was trained to capture the multi-modal longitudinal information from patient visit records (CXR images + reports) via a cross-attention-based multi-modal fusion module and a hierarchical memory-driven decoder. In contrast to previous works that only uses current visit data as input to train a model, our work exploits the longitudinal information available to pre-fill the "findings" section of radiology reports. Experiments show that our approach outperforms several recent approaches by $\geq$3% on F1 score, and $\geq$2% for BLEU-4, METEOR and ROUGE-L respectively. Code will be published at https://github.com/CelestialShine/Longitudinal-Chest-X-Ray.

Keywords: Chest X-Rays · Radiology reports · Longitudinal data · Report Pre-Filling · Report Generation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_19.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 189–198, 2023.
https://doi.org/10.1007/978-3-031-43904-9_19

1 Introduction

In current radiology practice, a signed report is often the primary form of communication, to communicate results of a radiological imaging exam between radiologist. Speech recognition software (SRS), which converts dictated words or sentences into text in a report, is widely used by radiologists. Despite SRS reducing the turn-around times for radiology reports, correcting any transcription errors in the report has been assumed by the radiologists themselves. But, persistent report communication errors due to SRS can significantly impact report interpretation, and also have dire consequences for radiologists in terms of medical malpractice [1]. These errors are most common for cross-sectional imaging exams (e.g., CT, MR) and chest radiography [2]. Problems also arise when re-examining the results from external examinations and in interventional radiology procedural reports. Such errors are due to many factors, including SRS finding a nearest match for a dictated word, the lack of natural language processing (NLP) for real-time recognition and dictation conversion [2], and unnoticed typographical mistakes. To mitigate these errors, a promising alternative is to automate the pre-filling of a radiology report with salient information for a radiologist to review. This enables standardized reporting via structured reporting.

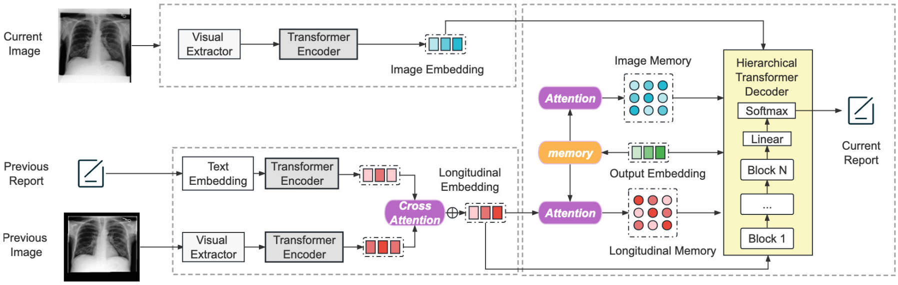


Fig. 1. Our proposed approach uses the CXR image and report from a previous patient visit and the current visit CXR image to pre-fill the "findings" section of the current visit report. The transformer-based model uses a cross-attention-based multi-modal fusion module and a hierarchical memory-driven decoder to generate the required text.

A number of methods to generate radiology reports have been proposed previously, with significant focus on CXR images [3–11]. Various attention mechanisms were proposed [4,6,12] to drive the encoder and the decoder to emphasize more informative words in the report, or visual regions in the CXR, and improve generation accuracy. Other approaches [8–10] effectively used Transformer-based models with memory matricies to store salient information for enhanced report generation quality. Despite these advances, there has been scarce research into harnessing *the potential of longitudinal patient visits* for improved patient care.

In practice, CXR images from multiple patient visits are usually examined simultaneously to find interval changes; e.g., a radiologist may compare a patient's current CXR to a previous CXR, and identify deterioration or improvement in the lungs for pneumonia. Reports from longitudinal visits contain valuable information regarding the patient's history, and harnessing the longitudinal multi-modal data is vital for the automated pre-filling of a comprehensive "findings" section in the report.

In this work, we propose to use longitudinal multi-modal data, i.e., previous visit CXR, current visit CXR, and previous visit report, to pre-fill the "findings" section of the patient's current visit report. To do so, we first gathered the longitudinal visit information for 26,625 patients from the MIMIC-CXR dataset[1] and created a new dataset called *Longitudinal-MIMIC*. Using this new dataset, we trained a transformer-based model containing a cross-attention-based multi-modal fusion module and a hierarchical memory-driven decoder to capture the features of longitudinal multi-modal data (CXR images + reports). In contrast to current approaches that only use the current visit data as input, our model exploits the longitudinal information available to pre-fill the "findings" section of reports with accurate content. Experiments conducted with the proposed dataset and model validate the utility of our proposed approach. Our main contribution in this work is training a transformer-based model that fully tackles the longitudinal multi-modal patient visit data to pre-fill the "findings" section of reports.

2 Methods

Dataset. The construction of the Longitudinal-MIMIC dataset involved several steps, starting with the MIMIC-CXR dataset, which is a large publicly available dataset of 377,110 chest X-ray images corresponding to 227,835 radiographic reports from 65,379 patients [13]. The first step in creating the Longitudinal-MIMIC dataset was to pre-process MIMIC-CXR to ensure consistency with prior works [8,9]. Specifically, patient visits where the report did not contain a "findings" section were excluded. For each patient visit, there was at least one chest X-ray image (frontal, lateral or other view) and a corresponding medical report. In our work, we only generated pre-filled reports with the "findings" section.

Table 1. A breakdown of the MIMIC-CXR dataset to show the number of patients with a specific number of visit records.

# visit records	1	2	3	4	5	>5
# patients	33,922	10,490	5,079	3,021	1,968	6,067

Next, the pre-processed dataset was partitioned into training, validation, and test sets using the official split provided with the MIMIC-CXR dataset. Table 1

[1] https://physionet.org/content/mimic-cxr-jpg/2.0.0/.

shows that 26,625 patients in MIMIC-CXR had ≥ 2 visit records, providing a large cohort of patients with longitudinal study data that could be used for our goal of pre-filling radiology reports. For patients with $\geq$2 visits, consecutive pairs of visits were used to capture richer longitudinal information. The dataset was then arranged chronologically based on the "StudyTime" attribute present in the MIMIC-CXR dataset. "StudyTime" represents the exact time at which a particular chest X-ray image and its corresponding medical report were acquired.

Following this, patients with $\geq$2 visit records were selected, resulting in 26,625 patients in the final *Longitudinal-MIMIC* dataset with a total of 94,169 samples. Each sample used during model training consisted of the current visit CXR, current visit report, previous visit CXR, and the previous visit report. The final dataset was divided into training (26,156 patients and 92,374 samples), validation (203 patients and 737 samples), and test (266 patients and 2,058 samples) splits. We aimed to create the *Longitudinal-MIMIC* dataset to enable the development and evaluation of models leveraging multi-modal data (CXR + reports) from longitudinal patient visits.

Model Architecture. Figure 1 shows the pipeline to generate a pre-filled "findings" section in the current visit report R_C, given the current visit CXR image I_C, previous visit CXR image I_P, and the previous visit report R_P. Mathematically, we can write: $p(R_C \mid I_C, I_P, R_P) = \prod_{t=1} p\left(w_t \mid w_1, \ldots, w_{t-1}, I_C, I_P, R_P\right)$, where w_i is the i-th word in the current report.

Encoder. Our model uses an *Image Encoder* and a *Text Encoder* to process the CXR images and text input separately. Both encoders were based on transformers. First, a pre-trained ResNet-101 [14] extracted image features $F = [f_1, \ldots, f_S]$ from the CXR images, where S is the number of patch features. They were then passed to the *Image Encoder*, which consisted of a stack of blocks. The encoded output was a list of encoded hidden states $H = [h_1, \ldots, h_S]$. The CXR images from the previous and the current visits were encoded in the same manner, and denoted by H^{I_P} and H^{I_C} respectively.

The *Text Encoder* encoded text information for language feature embedding using a previously published method [15]. First, the radiology report R_P was tokenized into a sequence of M tokens, and then transformed into vector representations $V = [v_1, \ldots, v_M]$ using a lookup table [16]. They were then fed to the *text encoder*, which had the same architecture as the *image encoder*, but with distinct network parameters. The final text feature embedding H^{R_P} was defined as: $H^{R_P} = \theta_R^E(V)$, where θ_R^E refers to the parameters of the report text encoder.

Cross-Attention Fusion Module. A multi-modal fusion module integrated longitudinal representations of images and texts using a cross-attention mechanism [17], which was defined as: $H^{I_P^*} = \operatorname{softmax}\left(\frac{q\left(H^{I_P}\right)k\left(H^{R_P}\right)^\top}{\sqrt{d_k}}\right) v\left(H^{R_P}\right)$ and $H^{R_P^*} = softmax\left(\frac{q\left(H^{R_P}\right)k\left(H^{I_P}\right)^\top}{\sqrt{d_k}}\right) v\left(H^{I_P}\right)$, where $q(\cdot), k(\cdot)$, and $v(\cdot)$ are linear transformation layers applied to features of proposals. d_k is the number of attention heads for normalization. Finally, $H^{I_P^*}$ and $H^{R_P^*}$ were concatenated to obtain the multi-modal longitudinal representations H^L.

Hierarchical Decoder with Memory. Our model's backbone decoder is a Transformer decoder with multiple blocks (The architecture of an example block is shown in the supplementary material). The first block takes partial output embedding H^O as input during training and a pre-determined starting symbol during testing. Subsequent blocks use the output from the previous block as input. To incorporate the encoded H^L and H^{I_C}, we use a hierarchical structure for each block that divides it into two sub-blocks: D^I and D^L.

Sub-block-1 uses H^{I_C} and consists of a self-attention layer, an encoder-decoder attention layer, and feed-forward layers. It also employs residual connections and conditional layer normalization [8]. The encoder-decoder attention layer performs multi-head attention over H^{I_C}. It also uses a memory matrix M to store output and important pattern information. The memory representations not only store the information of generated current reports over time in the decoder, but also the information across different encoders. Following [8], we adopted a matrix M to store the output over multiple generation steps and record important pattern information. Then we enhance M by aligning it with H^{I_C} to create an attention-aligned memory M^{I_C} matrix. Different from [8], we use M^{I_C} while transforming the normalized data instead of M. The decoding process of sub-block-1 D^I is formalized as: $H^{dec,b,I} = D^I(H^O, H^{I_C}, M^{I_C})$, where b stands for the block index. The output of sub-block 1 is combined with H^O through a fusion layer: $H^{dec,b} = (1-\beta)H^O + \beta H^{dec,b,I}$. β is a hyper-parameter to balance H^O and $H^{dec,b,I}$. In our experiment, we set it to 0.2.

The input to sub-block-2 D^L is $H^{dec,b}$. This structure is similar to sub-block-1, but interacts with H^L instead of H^{I_C}. The output of this block is $H^{dec,b,L}$ and combined with $H^{dec,b,I}$ by adding them together. After fusing these embeddings and doing traditional layer normalization for them, we use these embeddings as the output of a block. The output of the previous block is used as the input of the next block. After N blocks, the final hidden states are obtained and used with a Linear and Softmax layer to get the target report probability distributions.

3 Experiments and Results

Baseline Comparisons. We compared our proposed method against prior image captioning and medical report generation works respectively. The same *Longitudinal-MIMIC* dataset was used to train all baseline models, such as AoANet [18], CNNTrans [16], Transformer [15], R2gen [8], and R2CMN [9]. Implementation of these methods is detailed in the supplementary material.

Evaluation Metrics. Conventional natural language generation (NLG) metrics, such as $BLEU$ [19], $METEOR$ [20], and $Rouge_L$ [21] were used to evaluate the utility of our approach against other baseline methods. Similar to prior work [8,16], the CheXpert labeler [22] classified the predicted report for the presence

Table 2. Results of the NLG metrics (BLEU (BL), Meteor (M), Rouge R_L) and clinical efficacy (CE) metrics (Accuracy, Precision, Recall and F-1 score) on the *Longitudinal-MIMIC* dataset. Best results are highlighted in bold.

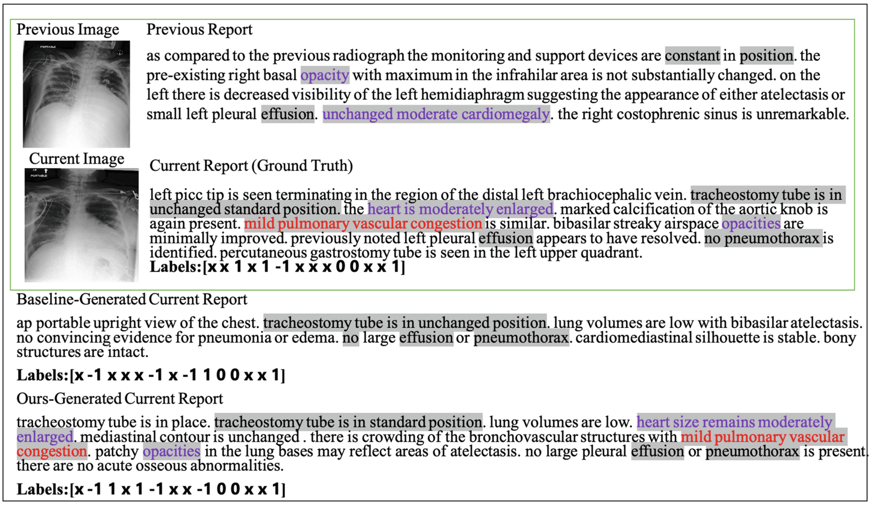
Method	NLG metrics						CE metrics			
	BL-1	BL-2	BL-3	BL-4	M	R_L	A	P	R	F-1
AoANet	0.272	0.168	0.112	0.080	0.115	0.249	0.798	0.437	0.249	0.317
CNN+Trans	0.299	0.186	0.124	0.088	0.120	0.263	0.799	0.445	0.258	0.326
Transformer	0.294	0.178	0.119	0.085	0.123	0.256	0.811	0.500	0.320	0.390
R2gen	0.302	0.183	0.122	0.087	0.124	0.259	0.812	0.500	0.305	0.379
R2CMN	0.305	0.184	0.122	0.085	0.126	0.265	0.817	0.521	0.396	0.449
Ours	**0.343**	**0.210**	**0.140**	**0.099**	**0.137**	**0.271**	**0.823**	**0.538**	**0.434**	**0.480**
Baseline	0.294	0.178	0.119	0.085	0.123	0.256	0.811	0.500	0.320	0.390
+ report	0.333	0.201	0.133	0.094	0.135	0.268	0.823	0.539	0.411	0.466
+ image	0.320	0.195	0.130	0.092	0.130	0.268	0.817	0.522	0.34	0.412
simple fusion	0.317	0.193	0.128	0.090	0.130	0.266	0.818	0.521	0.396	0.450

of 14 disease conditions[2] and compared them against the labels of the ground-truth report. Clinical Efficacy (CE) metrics, such as; accuracy, precision, recall, and F-1 score, were used to evaluate model performance.

Results. Table 2 shows the summary of the NLG metrics and CE metrics for the 14 disease observations for our proposed approach when compared against prior baseline approaches. In particular, our model achieves the best performance over previous baselines across all NLG and CE metrics.

Generic image captioning approaches like AoANet resulted in unsatisfactory performance on the *Longitudinal-MIMIC* dataset as they failed to capture specific disease observations. Moreover, our approach outperforms previous report generation methods, R2Gen and R2CMN that also use memory-based models, due to the added longitudinal context arising from the use of longitudinal multi-modal study data (CXR images + reports). In our results, the BLEU scores show a substantial improvement, particularly in BLEU-4, where we achieve a 1.4% increase compared to the previous method R2CMN. BLEU scores measure how many continuous sequences of words appear in predicted reports, while $Rouge_L$ evaluates the fluency and sufficiency of predicted reports. The highest $Rouge_L$ score demonstrates the ability of our approach to generate accurate reports, rather than meaningless word combinations. We also use METEOR for evaluation, taking into account the precision, recall, and alignment of words and phrases in generated reports and the ground truth. Our METEOR score shows a 1.1% improvement over the previous outstanding method, which further solidifies the effectiveness of our approach. Meanwhile, our model exhibits a significant

[2] No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Lesion, Airspace Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture and Support Devices.



Previous Report

as compared to the previous radiograph the monitoring and support devices are constant in position. the pre-existing right basal opacity with maximum in the infrahilar area is not substantially changed. on the left there is decreased visibility of the left hemidiaphragm suggesting the appearance of either atelectasis or small left pleural effusion. unchanged moderate cardiomegaly. the right costophrenic sinus is unremarkable.

Current Report (Ground Truth)

left picc tip is seen terminating in the region of the distal left brachiocephalic vein. tracheostomy tube is in unchanged standard position. the heart is moderately enlarged. marked calcification of the aortic knob is again present. mild pulmonary vascular congestion is similar. bibasilar streaky airspace opacities are minimally improved. previously noted left pleural effusion appears to have resolved. no pneumothorax is identified. percutaneous gastrostomy tube is seen in the left upper quadrant.

Labels:[x x 1 x 1 -1 x x x 0 0 x x 1]

Baseline-Generated Current Report

ap portable upright view of the chest. tracheostomy tube is in unchanged position. lung volumes are low with bibasilar atelectasis. no convincing evidence for pneumonia or edema. no large effusion or pneumothorax. cardiomediastinal silhouette is stable. bony structures are intact.

Labels:[x -1 x x x -1 x -1 1 0 0 x x 1]

Ours-Generated Current Report

tracheostomy tube is in place. tracheostomy tube is in standard position. lung volumes are low. heart size remains moderately enlarged. mediastinal contour is unchanged . there is crowding of the bronchovascular structures with mild pulmonary vascular congestion. patchy opacities in the lung bases may reflect areas of atelectasis. no large pleural effusion or pneumothorax is present. there are no acute osseous abnormalities.

Labels:[x -1 1 x 1 -1 x x -1 0 0 x x 1]

Fig. 2. Two examples of pre-filled "findings" sections of reports. Gray highlighted text indicates the same words or words with similar meaning that appear in the current reports and other reports. Purple highlighted text represents similar words in the current visit report generated by our approach, previous visit reports, and groundtruth current visit report. The red highlighted text indicates similar words that only exist in the report generated by our approach and the current ground truth report. R2Gen was the baseline method that generated the report. The "Labels" array shows the CheXpert classification of 14 disease observations (see text for details) as positive (1), negative (−1), uncertain (0) or unmentioned (×). (Color figure online)

improvement in clinical efficacy metrics compared to other baselines. Notably, F1 is the most important metric, as it provides a balanced measure of both precision and recall. Our approach outperforms the best-performing method by 3.1% in terms of F1 score. These observations are particularly significant, as higher NLG scores do not necessarily correspond to higher clinical scores [8], confirming the effectiveness of our proposed method.

Effect of Model Components. We also studied the contribution of different model components and detail results in Table 2. The *Baseline* experiment refers to a basic Transformer model trained to generate a pre-filled report given a chest CXR image without any additional longitudinal information. The NLG and CE metrics are poor for the vanilla transformer compared to our proposed approach. We also analyze the contributions of the previous chest CXR image + *image* and previous visit report + *report* when added to the model separately. These two experiments included memory-enhanced conditional normalization. We observed that with each added feature enhanced the pre-filled report quality compared to the baseline, but the previous visit report had a higher impact than the previous CXR image. We hypothesize that the previous visit reports contain more text that can be directly transferred to the current visit reports.

In our *simple fusion* experiment, we removed the cross-attention module and concatenated the encoded embeddings of the previous CXR image and previous visit report as one longitudinal embedding, while retaining the rest of the model. We saw a performance drop compared to our approach on our dataset, and also noticed that the results were worse than using the images or reports alone. These experiments demonstrate the utility of the cross-attention module in our proposed work.

4 Discussion and Conclusion

Case Study. We also ran a qualitative evaluation of our proposed approach on two cases as seen in Fig. 2. In these cases, we compare our generated report with the report generated by the R2Gen. In the first case, certain highlighted words in purple, such as "status post", "aortic valve" and "cardiac silhouette in the predicted current visit report are also seen in the previous visit report. The CheXpert classified "Labels" also show the pre-filled "findings" generated is highly consistent with the ground truth report in contrast to the R2Gen model. For example, the "cardiac silhouette enlarged" was not generated by the R2Gen model, but our prediction contains them and is consistent with the word "cardiomegaly" in the ground truth report. In the second case, our generated report is also superior. Not only does our report generate more of the same content as the ground truth, but the positive diagnosis labels classified by CheXpert in our report are completely consistent with those in the ground truth. We also provide more cases in the supplementary material.

Error Analysis. To analyze errors from our model, we examine generated reports alongside ground truths and longitudinal information. It is found that the label accuracy of the observations in the generated reports is greatly affected by the previous information. For example, as time changes, for the same observation "pneumothorax", the label can change from "positive" to "negative". And such changing examples are more difficult to generate accurately. According to our statistics, on the one hand, when the label results of current and previous report are the same, 88.96% percent of the generated results match them. On the other hand, despite mentioning the same observations, when the labels of current and previous report are different, there is an 84.42% probability of generated results being incorrect. Thus how to track and generate the label accurately of these examples is a possible future work to improve the generated radiology reports. One possible way to address this issue is to use active learning [23] or curriculum learning [24] methods to differentiate different types of samples and better train the machine learning models.

Conclusion. In this paper, we propose to pre-fill the "findings" section of chest X-Ray radiology reports by considering the longitudinal multi-modal (CXR images + reports) information available in the MIMIC-CXR dataset. We gathered 26,625 patients with multiple visits to constitute the new *Longitudinal-MIMIC* dataset, and proposed a model to fuse encoded embeddings of multi-modal data along with a hierarchical memory-driven decoder. The model generated a pre-filled "findings" section of the report, and we evaluated the generated results against prior image captioning and medical report generation works. Our model yielded a $\geq 3\%$ improvement in terms of the clinical efficacy F-1 score on the *Longitudinal-MIMIC* dataset. Moreover, experiments that evaluated the utility of different components of our model proved its effectiveness for the task of pre-filling the "findings" section of the report.

Acknowledgements. This research was supported by the Intramural Research Program of the National Library of Medicine and Clinical Center at the NIH. The authors thank to Qingyu Chen and Xiuying Chen for their time and effort in providing thoughtful comments and suggestions to revise this paper. This work was also supported by the National Institutes of Health under Award No. 4R00LM013001 (Peng), NSF CAREER Award No. 2145640 (Peng), and Amazon Research Award (Peng).

References

1. Smith, J.J., Berlin, L.: Signing a colleague's radiology report. Am. J. Roentgenol. **176**(1), 27–30 (2001). PMID: 11133532
2. Ringler, M.D., Goss, B.C., Bartholmai, B.J.: Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Inform. J. **23**(1), 3–13 (2017)
3. Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2506 (2016)
4. Jing, B., Xie, P., Xing, E.: 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), pp. 2577–2586 (2018)
5. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
6. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9049–9058 (2018)
7. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: on exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274 (2020)
8. Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1439–1449 (2020)
9. Teo, T.W., Choy, B.H.: STEM education in Singapore. In: Tan, O.S., Low, E.L., Tay, E.G., Yan, Y.K. (eds.) Singapore Math and Science Education Innovation. ETLPPSIP, vol. 1, pp. 43–59. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1357-9_3
10. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision. ECCV 2022. LNCS, vol. 13695, pp. 563–579. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19833-5_33

11. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13753–13762 (2021)
12. Xue, Y., et al.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2018)
13. Johnson, A.E.W., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Moon, J.H., Lee, H., Shin, W., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. arXiv preprint arXiv:2105.11333 (2021)
17. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Adv. Neural. Inf. Process. Syst. **34**, 14200–14213 (2021)
18. Huang, L., Wang, W., Chen, J., Wei, X.-Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: ACL 2002, pp. 311–318 (2002)
20. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evasukhbaatar2015endluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
21. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
22. Irvin, J., et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc. AAAI Conf. Artif. Intell. **33**, 590–597 (2019)
23. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, pp. 1–114. Springer, Cham (2012). https://doi.org/10.1007/978-3-031-01560-1
24. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48 (2009)