

A Self-training Framework for Automated Medical Report Generation

Siyuan Wang^{1*} and Zheng Liu^{2*†} and Bo Peng³

¹The University of Sydney, Australia

²University College London, UK

³Newcastle University, UK

Abstract

Medical report generation, focusing on automatically generating accurate clinical findings from medical images, is an important medical artificial intelligence task. It reduces the workload of physicians in writing reports. Many of the current methods depend heavily on labeled datasets that include image-report pairs, but such datasets labeled by physicians are hard to acquire in clinical practice. In this paper, we introduce a self-training framework named REMOTE (i.e., Revisiting sELf-training for Medical repOrT gENERation) to exploit the unlabeled medical images and a MedCLIPScore to augment a small-scale dataset for training the medical report generation model. Experiments conducted on the MIMIC-CXR benchmark dataset and a COVID-19 dataset demonstrate that, our REMOTE framework, using only 1% labeled training data, achieves competitive performance with previous methods that are trained on entire training data.

1 Introduction

Generating medical reports automatically involves producing clinical descriptions based on the input visual medical images (Jing et al., 2018, 2019; Li et al., 2018; Liu et al., 2021b). This is similar to the task of image captioning (Xu et al., 2015; Chen et al., 2015), which aims to generate visual descriptions to describe the input images. Therefore, based on the benchmark dataset MIMIC-CXR (Johnson et al., 2019), inspired by the success of image captioning, various state-of-the-art data-driven models, especially those based on the encoder-decoder structure (Chen et al., 2020; Liu et al., 2021b; Wang et al., 2022a), have achieved significant advancements. However, medical data labeling requires specialized expertise from physicians and also involves privacy concerns. Therefore, acquiring medical report generation datasets is time-consuming

and costly (Liu et al., 2021c). As a result, when compared to datasets used for general image captioning datasets such as Conceptual Captions (Soricut et al., 2018), the size of the medical dataset MIMIC-CXR is relatively small. This size limitation becomes a challenge when dealing with novel diseases like COVID-19, where collecting and labeling adequate training data promptly is difficult. It hinders the application of existing medical report generation models in addressing novel diseases to alleviate the workload of physicians efficiently.

Considering that there are a lot of public image-only datasets, e.g., CheXpert (Irvin et al., 2019), RSNA Pneumonia (Shih et al., 2019), COVID images (Rahman et al., 2021), in the literature. To this end, we propose a self-training framework REMOTE, which enhances the performance of the medical report generation model by simultaneously utilizing high-quality paired image-report datasets and image-only datasets. In implementation, we adopt the Noisy Student self-training framework (Xie et al., 2020; He et al., 2020) as the basis to build our REMOTE for medical report generation, which consists of a “teacher” model and a “student” model. It begins by training a “teacher” model on the high-quality annotated image-report pairs, e.g., MIMIC-CXR dataset (Johnson et al., 2019). Subsequently, the teacher model is used to generate pseudo-reports for medical images in the image-only dataset without annotated reports. We then employ MedCLIPScore to score each generated pseudo image-report pair and filter out low-scoring pseudo image-report pairs. Finally, we train a “student” model on both the annotated high-quality image-report pairs and the generated pseudo image-report pairs. In the next training step, we consider the “student” model as the new “teacher” model, and by repeating the above steps, we can generate new pseudo image-report pairs and train new “student” models. Through iterating these steps, we ultimately obtain an accurate and robust medical

*Equal Contributions.

†Corresponding author: zhengliu.ucl@gmail.com.

report generation model.

It is worth noting that, while the self-training framework has been explored in uni-modal tasks such as image classification (Xie et al., 2020) and machine translation (He et al., 2020), self-training in medical report generation has not been well explored. This is because medical report generation is a multi-modal medical task, incorporating disparities between the visual and the textual modalities. Thus, inspired by the great success of CLIP (Radford et al., 2021), which is trained to align image and text modalities, we follow the CLIPScore (Hessel et al., 2021) to construct the MedCLIPScore to obtain a high-quality pseudo image-report pairs. In detail, we train the MedCLIP (Wang et al., 2022b) on the MIMIC-CXR dataset, and use it as MedCLIPScore to boost the performance and robustness of the medical report generation model.

Overall, the main contributions of this paper are as follows:

- Based on the noisy student self-training framework, we propose a self-training framework REMOTE for automated medical report generation with limited labeled training data.
- Our proposed method includes three components: “teacher” model, MedCLIPScore, and “student” model. The “teacher” model and MedCLIPScore focus on obtaining high-quality pseudo image-report pairs from image-only datasets, which are used to obtain a robust “student” model. By taking the “student” model as the new “teacher” model and iterating the above steps, REMOTE can achieve strong performances with limited labeled data.
- Experiments on two datasets show that our method can achieve competitive results with existing fully-supervised methods with only 1% labeled training data.

2 Approach

In this section, we will introduce the core three components of our approach, i.e., the “teacher” model, MedCLIPScore, and the “student” model.

2.1 Formulation

The goal of medical report generation is to generate an accurate medical report r given the input medical image i (Jing et al., 2019; Liu et al., 2021c), which can be formulated as:

$$\text{Medical Report Generation} : i \rightarrow r, \quad (1)$$

The encoder-decoder framework has been widely used in medical report generation, in which the image encoder is designed to extract image embeddings of input medical image i and the text decoder is designed to generate the target report r .

2.2 “Teacher” Model

In this study, we employ the ResNet-50 (He et al., 2016) and Transformer (Vaswani et al., 2017) to implement the image encoder and the text decoder of the “teacher” model, respectively. To train the model, we adopt the high-quality and widely-used medical report generation dataset MIMIC-CXR dataset (Johnson et al., 2019). Given the ground truth report $r = \{y_1, y_2, \dots, y_N\}$ labeled by physicians for the input image i , we can train the model by minimizing the cross-entropy (CE) loss, defined as follows:

$$L_{\text{CE}}(\theta) = - \sum_{i=1}^N \log(p_{\theta}(y_i | y_{1:i-1}; i)) \quad (2)$$

However, image-report pairs labeled by physicians are hard to obtain in the real world. In this study, we propose to adopt the medical images without labeled reports from the image-only datasets to boost the performance.

2.3 MedCLIPScore

We first adopt the trained “teacher” model to generate the reports for the medical images from the image-only dataset, e.g., CheXpert (Irvin et al., 2019). Given the input image i^* , we adopt the beam-search decoding with beam size B to generate B medical reports $\{r_1^*, r_2^*, \dots, r_B^*\}$. In this way, we can obtain B pseudo pairs of image and report, i.e., $\{(i^*, r_1^*), (i^*, r_2^*), \dots, (i^*, r_B^*)\}$.

Next, to obtain high-quality pseudo image-report pairs, MedCLIPScore is introduced. We further pre-train the MedCLIP (Wang et al., 2022b; Zhang et al., 2020) on the MIMIC-CXR dataset (Johnson et al., 2019) using Image-Text Matching (ITM) pre-training objective, which aims to distinguish whether an image-report pair is a match. In detail, positive image-report pairs and randomly sampled negative pairs are fed into the MedCLIP and the concatenation of textual representation of report R and visual representation of image I is processed by a softmax layer to output a binary probability p_{ITM} . Therefore, the ITM objective is defined as:

$$\mathcal{L}_{\text{ITM}} = - \sum_{(I,R)} \log p_{\text{ITM}}(Y_{\text{ITM}} | I, R) \quad (3)$$

Methods	Ratio of Labeled Data	MIMIC-CXR				COVID-19			
		BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
AdaAtt (Lu et al., 2017)	100%	0.088	0.118	0.266	0.084	0.054	0.084	0.136	0.061
BUTD (Anderson et al., 2018)	100%	0.074	-	0.250	0.073	0.066	0.102	0.159	0.060
Trans. (Vaswani et al., 2017)	100%	0.090	0.125	0.265	-	0.071	0.105	0.176	0.069
R2Gen (Chen et al., 2020)	100%	0.103	0.142	0.277	-	0.078	0.114	0.198	0.135
PPKED (Liu et al., 2021b)	100%	0.106	0.149	0.284	0.237	-	-	-	-
DeltaNet (Wu et al., 2022)	100%	0.114	-	0.277	0.281	-	-	-	-
XProNet (Wang et al., 2022a)	100%	0.105	0.138	0.279	-	0.088	0.120	0.213	0.152
REMOTE	1%	0.115	0.147	0.289	0.266	0.102	0.135	0.254	0.201
REMOTE*	100%	0.125	0.157	0.304	0.338	0.113	0.146	0.278	0.269
REMOTE*	100%	0.182	0.265	0.412	0.497	0.203	0.266	0.391	0.407

Table 1: Results of our proposed approach on two datasets. Existing works are trained on 100% labeled image-report pairs of the downstream dataset. 1% denotes our method is trained on 1% labeled image-report pairs plus 90% remaining image-only data (without reports) of the downstream dataset. We calculate P-values between the performances of our REMOTE (100%) and the best-performing baseline. * denotes we adopt the external public unlabeled image-only datasets for training.

After obtaining the cross-modal model MedCLIP pre-trained on MIMIC-CXR, we follow CLIPScore (Hessel et al., 2021) to obtain MedCLIPScore, which can measure the match between the image and pseudo report. Given a pseudo image-report pair (i^*, r^*) , we first adopt the MedCLIP to extract the visual representation I^* and textual representation R^* . Then, to assess the quality of the generated pseudo image-report pair (i^*, r^*) , following CLIPScore (Hessel et al., 2021), MedCLIPScore is defined as:

$$\text{MedCLIPScore}(I^*, R^*) = w * \max(\langle I^*, R^* \rangle, 0) \quad (4)$$

where w attempts to stretch the range of the score distribution to $[0, 1]$ (Hessel et al., 2021), and the $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. At last, we set a threshold τ and filter out low-scoring pseudo image-report pairs with MedCLIPScore lower than τ . In this way, we can obtain high-quality pseudo image-report pairs to boost the medical report generation model.

2.4 “Student” Model

We further combine the generated pseudo image-report pairs and original human-labeled image-report pairs to train the “student” model by minimizing the cross-entropy loss. During the training process, for each training batch, the ratio of human-labeled and generated pseudo image-report pairs is 1: M ($M=6$). At last, after obtaining the “student” model, we take the trained “student” model as the new “teacher” model to generate new pseudo image-report pairs. The REMOTE stops training until the performance no longer increases. As a result, REMOTE can achieve encouraging results with limited labeled data.

3 Experiments

We first describe a benchmark dataset, our built dataset, the metrics, and the settings used for evaluation. Then, we present the results of our approach.

3.1 Datasets, Metrics, and Settings

3.1.1 Datasets

We conduct experiments on the widely-used benchmark dataset MIMIC-CXR (Johnson et al., 2019) and a built COVID-19 dataset, where the former consists of 377,110 chest X-ray images and 227,835 radiology reports, the latter includes 2,025 COVID-19 cases (including both images and reports). For the MIMIC-CXR dataset, we follow the official splits to report our results. The COVID-19 dataset contains 980 COVID-19 records and 1,045 non-COVID-19 records from 1,877 patients, with a total of 2,025 records. Each record is composed of the X-ray image and the corresponding medical report. The max, median, and mean length of the reports are 82, 35, and 39 words, respectively. We randomly split the dataset into 70%-10%-20% training-validation-testing splits.

3.1.2 Metrics

To access the performance of medical report generation models, we compute the widely-used evaluation metrics, i.e., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), which measure the match between the generated reports and ground truth reports.

3.1.3 Settings

To implement our approach and the existing works on the COVID-19 dataset, we follow common practice (Jing et al., 2019; Li et al., 2018; Liu et al.,

Methods	Ratio of Labeled Data	MIMIC-CXR				COVID-19			
		BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
XProNet (Wang et al., 2022a)	100%	0.105	0.138	0.279	-	0.088	0.120	0.213	0.152
REMOTE (Ours)	1%	0.115	0.147	0.289	0.266	0.102	0.135	0.254	0.201
w/o input noise	1%	0.103	0.131	0.276	0.254	0.097	0.126	0.237	0.192
w/o model noise	1%	0.093	0.125	0.270	0.248	0.093	0.114	0.228	0.178
REMOTE (Ours)	100%	0.125	0.157	0.304	0.338	0.113	0.146	0.278	0.269
w/o input noise	100%	0.119	0.152	0.295	0.329	0.102	0.137	0.271	0.260
w/o model noise	100%	0.108	0.140	0.282	0.314	0.095	0.129	0.266	0.251
REMOTE* (Ours)	100%	0.125	0.157	0.304	0.338	0.113	0.146	0.278	0.269
w/ RSNA Pneumonia (Shih et al., 2019)	100%	0.146	0.190	0.355	0.381	0.142	0.180	0.310	0.301
w/ CheXpert (Irvin et al., 2019)	100%	0.165	0.231	0.381	0.422	0.172	0.215	0.349	0.353
w/ COVID images (Rahman et al., 2021)	100%	0.136	0.176	0.332	0.365	0.175	0.204	0.352	0.358
w/ RSNA + CheXpert + COVID	100%	0.182	0.265	0.412	0.497	0.203	0.266	0.391	0.407

Table 2: Quantitative analysis of our method, including the input noise and model noise. * denotes the model is further trained on the external unlabeled image-only data.

2021b,a) to adopt the ResNet-50 (He et al., 2016) as the image encoder, which is pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on public available CheXpert dataset (Irvin et al., 2019). We adopt the Transformer (Vaswani et al., 2017) to implement the text decoder. For the MedCLIPScore, we set w to 2.5, and threshold τ to 0.75 (Hessel et al., 2021). For model training, we follow the Noisy Student self-training framework (Xie et al., 2020) to inject the input noise and model noise to enhance the robustness and performance of the model. For parameters optimization, we adopt the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e-4$ and a batch size of 16. We apply a beam search of size 3 for inference.

3.2 Main Results

The results in Table 1 show the comparison of REMOTE and existing strong methods, e.g., PPKED (Liu et al., 2021b), DeltaNet (Wu et al., 2022), and XProNet (Wang et al., 2022a). As we can see, when trained on 1% of labeled image-report pairs, combined with 90% of image-only data without reports, REMOTE demonstrates competitive performance with previous methods. When REMOTE is trained with the complete 100% labeled data, its performance surpasses the existing state-of-the-art methods on both two datasets. For example, on the MIMIC-CXR dataset, REMOTE achieves a BLEU-4 score of 0.125, a METEOR score of 0.157, a ROUGE-L score of 0.304, and a CIDEr score of 0.338. Similarly, on the COVID-19 dataset, the scores are 0.113, 0.146, 0.278, and 0.269, respectively. Additionally, the table also indicates an enhanced version of REMOTE, denoted as REMOTE*, which utilizes external large-scale public unlabeled image-only datasets, i.e., RSNA Pneumonia (Shih et al., 2019), CheXpert (Irvin

et al., 2019), COVID chest X-ray images (Rahman et al., 2021; Cohen et al., 2020), for training. The results show a remarkable improvement. Overall, the result demonstrates the robustness and effectiveness of the REMOTE, especially when considering its competitive results with limited labeled training data, which are common in real-life clinical practice.

4 Analysis

We will provide quantitative and qualitative analyses to understand our method.

4.1 Quantitative Analysis

Table 2 shows that both the input noise and model noise contribute to improved performances. In detail, the input noise and model noise lead to the best improvements when the model is trained on the 1% labeled data. It proves the effectiveness of our approach in generating accurate medical reports when the labeled training data is limited. When adopting the external unlabeled image-only data for model training, the performances of our approach on the two datasets are improved with the increasing amount of unlabeled image-only data. As a result, our method greatly surpasses the state-of-the-art method XProNet (Wang et al., 2022a) by 13.3% ROUGE-L score and 25.5% CIDEr score on the MIMIC-CXR and COVID-19 datasets, respectively. The superior performances show that our approach has the potential to be well-applied to real clinical practice, where the training data labeled by physicians is scarce.

4.2 Qualitative Analysis

In this section, we show the medical reports generated to compare our approach with state-of-the-art method (Wang et al., 2022a) qualitatively. The


	<p>Ground Truth: The heart is again mildly enlarged. The mediastinal and hilar contours appear unchanged. Pleural effusions have more fully resolved. There is persistent patchy opacification of the right mid upper and left upper lungs, which are background findings. Streaky left basilar opacity also has improved. Pulmonary edema has more fully resolved. A PICC line again terminates in the superior vena cava.</p>	<p>XProNet [100%]: Lungs are clear and the pulmonary vasculature is normal. Heart size is normal. There is no pleural effusion or pneumothorax. There are no acute osseous abnormalities. The mediastinal and hilar contours are normal.</p>	<p>REMOTE [1%]: There is a moderate cardiomegaly with bilateral patchy opacities. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is seen.</p>
---	---	--	---

Figure 1: Medical reports generated by the state-of-the-art fully-supervised method XProNet (Wang et al., 2022a) and our approach. Correct results and Unfavorable results are denoted as Bold text and Underlined text, respectively.

generated reports are reported in Figure 1, which clearly shows that there are two key abnormalities in the given medical images, i.e., “{The heart is again mildly enlarged}” and “{There is persistent patchy opacification of the right mid upper and left upper lungs}”. Fortunately, REMOTE trained on 1% labeled training data correctly generate “{There is a moderate cardiomegaly with bilateral patchy opacities}”, which not only captures the ‘enlarged heart’ and ‘patchy opacification’, but also accurately describes the details, i.e., ‘bilateral’. In comparison, the report generated by the state-of-the-art method XProNet, which is trained on full training data, does not cover the two core abnormalities. Besides, XProNet gives a wrong clinical description, i.e., “{Heart size is normal}”. The reported example qualitatively demonstrates the effectiveness of REMOTE, which can generate accurate medical reports using 1% labeled training data.

5 Conclusion

In this work, we presented a novel self-training framework, REMOTE, aimed at boosting the performance of medical report generation from unlabeled visual medical images, especially in scenarios with limited labeled training data. By harnessing the power of both high-quality paired image-report datasets and unlabeled image-only datasets, our approach effectively reduces the reliance on extensive labeled training data, which are both time-consuming and costly to obtain. Our REMOTE introduces a “teacher” model and a “student” model to generate pseudo image-report pairs and refine the generation, respectively. Specifically, we introduced MedCLIPScore to ensure the quality of the generated pseudo image-report pairs. The experiments on the widely-used benchmark dataset MIMIC-CXR and a COVID-19 dataset validated

the robustness and effectiveness of REMOTE. In particular, using only 1% of labeled training data, our approach could achieve competitive performances comparable to fully-supervised state-of-the-art methods.

Limitations

Although the REMOTE can automatically generate medical report generation with limited labeled training data, the performance of our approach is highly dependent on the amount and quality of unlabeled image-only data. It indicates that we still need the independent set of medical images which may still be difficult to collect for some scenarios or some types of medical images. Besides, the iterative “teacher” and “student” training model can be computationally intensive, possibly limiting its adoption in resource-constrained settings.

The error analysis shows that our model suffers from several common drawbacks: i) generating repeated sentences, and ii) misunderstanding rare pathologies and diseases in some cases. They can be attributed to the lack of detailed and accurate visual information. We may alleviate these drawbacks by introducing strong pathologies and disease predictors to accurately extract a set of pathologies and diseases. However, it is unlikely to be avoided completely, as these drawbacks are common in medical report generation models.

Ethics Statement

We conduct experiments and analysis on a public MIMIC-CXR dataset and a built COVID-19 dataset. Besides the COVID-19 dataset, all medical images used, both labeled and unlabeled, were sourced from publicly available datasets, ensuring that no private or unauthorized patient data was employed. For the COVID-19 dataset, all protected

health information (PHI) was removed. Furthermore, all images and reports were de-identified, ensuring the privacy and confidentiality of patients. While our method reduces the need for extensive labeled datasets, its outputs are still machine-generated, requiring critical human oversight, particularly when used in clinical decision-making.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. In *CVPR*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEValuation@ACL*.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *EMNLP*.
- Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q. Duong, and Marzyeh Ghassemi. 2020. COVID-19 image data collection: Prospective predictions are the future. *CoRR*, abs/2006.11988.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
- Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *ICLR*. OpenReview.net.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, pages 7514–7528. Association for Computational Linguistics.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpankaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*.
- Baoyu Jing, Zeya Wang, and Eric P. Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *ACL*.
- Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *ACL*.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NeurIPS*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.
- Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *ACL*.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*.
- Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. 2021c. Auto-encoding knowledge graph for unsupervised medical report generation. In *NeurIPS*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for automatic evaluation of machine translation. In *ACL*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.
- Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas M. Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al-Máadeed, Susu M. Zughaiier, Muhammad Salman Khan, and Muhammad Enamul Hoque Chowdhury. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest x-ray images. *Comput. Biol. Medicine*, 132:104319.

- George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence*, 1(1).
- Radu Soricut, Nan Ding, Piyush Sharma, and Sebastian Goodman. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.
- Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *ECCV*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.
- Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S. Kevin Zhou, and Li Xiao. 2022. Deltanet: Conditional medical report generation for COVID-19 diagnosis. In *COLING*.
- Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10684–10695. Computer Vision Foundation / IEEE.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.