PhenotypeCLIP: Phenotype-based Contrastive Learning for Medical Imaging Report Generation

Siyuan Wang[1]* and **Bo Peng**[2]* and **Yichao Liu**[2] and **Qi Peng**[2]
[1]The University of Sydney, Australia
[2]Newcastle University, UK
mariobopeng.work@gmail.com

Abstract

Given the input radiology images, the objective of medical report generation is to produce accurate and comprehensive medical reports, which typically include multiple descriptive clinical sentences associated with different phenotypes. Most existing approaches have relied on a pretrained vision encoder to extract the visual representations of the images. In this study, we propose a phenotype-based contrastive learning framework, i.e., PhenotypeCLIP, to efficiently bridge the gap between visual and textual modalities for improved text generation. In contrast to existing contrastive learning methods which learn representations by contrasting images with entire reports, our approach learns more fine-grained representations, i.e., phenotype-based representations, by contrasting images with each sentence within the reports. The experiments on two widely-used datasets MIMIC-CXR and IU X-ray demonstrate that PhenotypeCLIP can achieve promising performances and substantially outperform the conventional contrastive learning methods.

1 Introduction

Medical images, such as those from radiology and pathology, and medical reports, which consist of multiple clinical sentences describing both the normalities and abnormalities in the medical images, are frequently used for disease diagnosis and treatment (Jing et al., 2018, 2019). Therefore, medical report generation has the potential to reduce the heavy workload of experienced radiologists in report writing and remind inexperienced radiologists of abnormalities (Jing et al., 2018; Li et al., 2018).

Similar to the task of image captioning (Xu et al., 2015), which aims to describe the visual content in the images, lots of encoder-decoder-based medical report generation models are proposed (Jing et al., 2018, 2019; Li et al., 2018; Chen et al., 2020c; Liu et al., 2021a; Wang et al., 2022a). In the encoding stage, the visual representations of the images are extracted by a vision encoder pretrained on ImageNet (Deng et al., 2009); In the decoding stage, the medical report is generated by a Transformer (Chen et al., 2020c; Liu et al., 2021b) or LSTM (Jing et al., 2018). Specifically, several works propose to further fine-tune the vision encoder on medical image datasets, e.g., ChestX-ray8 (Wang et al., 2017) and CheXpert (Irvin et al., 2019). Nevertheless, we argue that as a text-oriented task, a core step of the medical report generation models is to efficiently bridge the gap between visual and textual modalities. To this end, in this work, motivated by the great success of contrastive learning in bridging the gap between visual and textual modalities, i.e., CLIP (Radford et al., 2021), we introduce a phenotype-based contrastive learning framework - PhenotypeCLIP.

In implementation, existing contrastive learning models first extract the visual representations of images and textual representations of entire reports, and then pre-train models by contrasting the visual representations with the textual representations. In this study, we (i) first construct a set of phenotypes; (ii) introduce an attention mechanism to transform each sentence within the reports into phenotype-based textual representations; (iii) adopt the phenotype-based textual representations to extract phenotype-based visual representations; (iv) and per-train the model by contrasting the two types of phenotype-based representations. In this way, by splitting the entire report into multiple sentences, PhenotypeCLIP not only learns fine-grained representations, but also scales up contrasting learning, resulting in boosting the downstream medical report generation task.

We perform the experiments on two benchmark datasets, i.e., MIMIC-CXR (Johnson et al., 2019) and IU X-ray (Demner-Fushman et al., 2016). The results validate the effectiveness of PhenotypeCLIP, which substantially outperforms the conventional

*Both authors contributed equally to this paper.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15949–15956
December 6-10, 2023 ©2023 Association for Computational Linguistics

contrastive learning methods on all widely used evaluation metrics.

Overall, the main contributions of this paper are:

- In this study, to efficiently bridge the gap between visual and textual modalities, we perform phenotype-based contrastive learning to learn accurate and fine-grained representations, which can boost the downstream text-oriented medical report generation task.

- The experiments and analysis performed on two benchmark datasets demonstrate that the proposal achieves improved performances on all metrics.

2 Related Work

We will introduce our related work from medical report generation and contrastive learning.

2.1 Medical Report Generation

Medical report generation aims to interpret the medical image by generating a report (Jing et al., 2018, 2019; Chen et al., 2020c). In contrast to image captioning (Xu et al., 2015; Anderson et al., 2018; Liu et al., 2018; Gu et al., 2022) that produces a single-sentence description for general images, the medical report generation aims to produce a paragraph containing multiple clinical descriptions.

Inspired by the success of image captioning, numerous encoder-decoder-based frameworks have been introduced for medical report generation. For example, Jing et al. (2018) and Chen et al. (2020c) respectively introduced a hierarchical LSTM with an attention mechanism and a Transformer to learn to generate the long paragraph; Yuan et al. (2019); Jing et al. (2018) and You et al. (2021) further incorporated the medical concepts to boost the performance; (Yang et al., 2022; Liu et al., 2021b) and Li et al. (2019) proposed to construct the medical knowledge graph to inject the medical knowledge into the models.

In summary, while deep learning models, particularly those encoder-decoder-based frameworks, have achieved promising results for medical report generation, they mainly adopt the vision encoder pre-trained on ImageNet (Deng et al., 2009) and CheXpert (Irvin et al., 2019) to extract the visual representations. In this study, we argue that as a text-oriented task, it is necessary to explicitly bridge the gap between visual and textual domains. To this end, we propose phenotype-based contrastive learning to learn fine-grained representation and thus boost the downstream task.

2.2 Contrastive Learning

In recent years, contrastive learning, which trains the models to distinguish between positive and negative pairs, has achieved state-of-the-art performances in visual representation learning (Chen et al., 2020a; He et al., 2020; Chen et al., 2020b) and vision-language representation learning (Radford et al., 2021). Inspired by the great success of contrastive learning, several works (Huang et al., 2021; Zhang et al., 2020a; Boecking et al., 2022; Zhou et al., 2023, 2022; Wang et al., 2022b) have been proposed to learn robust and accurate medical vision-language representations, which can be used to achieve promising results on various downstream tasks. However, most existing works focus on contrasting images with entire reports, ignoring the potential phenotypes in each sentence within the reports. To this end, we propose the phenotype-CLIP to implement phenotype-based contrastive learning to learn fine-grained representations, outperforming existing works.

3 Approach

In this section, we introduce the proposal for medical report generation in detail.

3.1 Formulation of Contrastive Learning

Contrastive learning has shown encouraging results in bridging the gap between vision and language domains, e.g., CLIP (Radford et al., 2021). Given a batch of N training samples, including N pairs of image and report, i.e., $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, where x_i denotes the visual representations of the i-th image and y_i denotes the textual representations of the i-th report. Therefore, (x_i, y_i) dentoes the positive sample, and (x_i, y_j), where $i \neq j$, denotes the negative sample. To train the models, contrastive learning adopts the InfoNCE loss, which maximizes the mutual information between x_i and y_i, and minimizes the mutual information between x_i and y_j, which can be defined as follows:

$$\begin{aligned}
\mathcal{L}^{x\rightarrow y} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{sim}(x_i, y_j)/\tau)} \\
\mathcal{L}^{y\rightarrow x} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{sim}(y_i, x_j)/\tau)} \qquad (1) \\
\mathcal{L}_{\mathrm{CL}} &= \frac{1}{2}\left(\mathcal{L}^{x\rightarrow y} + \mathcal{L}^{y\rightarrow x}\right),
\end{aligned}$$

where $\mathbf{sim}(\cdot)$ represents the cosine similarity and τ is a hyper-parameter.

3.2 PhenotypeCLIP

As we can see, although contrastive learning has been well explored for medical representation learning in existing literature (Huang et al., 2021; Zhang et al., 2020a; Boecking et al., 2022; Zhou et al., 2023, 2022; Wang et al., 2022b), they encode the entire report y_i for training.

In this study, we argue that each sentence within the report contains different phenotypes. Thus, we propose the PhenotypeCLIP to capture the phenotypes for better medical representation learning. For example, if a medical report contains (phenotype$_1$, phenotype$_2$) and another report contains (phenotype$_1$, phenotype$_3$), then existing works would treat the two reports as negative samples. However, we can notice that the two different reports contain the same phenotype, i.e., phenotype$_1$. Therefore, phenotype-based contrastive learning can enable the models to learn such fine-grained representations. In implementation:

(i) We first construct a set of phenotypes, $P = \{p_1, p_2, \dots, p_M\}$, where M stands for the total number of the phenotypes. $p_i \in \mathbb{R}^d$ can be either a randomly initialized soft vector or a word embedding of a pre-defined phenotype, e.g., cardiomegaly.

(ii) We introduce the three-layer Transformer (Vaswani et al., 2017), which includes a multi-head attention and a feed-forward network[1], to extract the phenotype-based representations. Given a pair of image and report (x_i, y_i), where report contains K sentences, i.e., $y_i = \{s_{i1}, s_{i2}, \dots, s_{iK}\}$. We first transform each sentence s_{ik} within the reports into phenotype-based textual representations $s^P_{ik} \in \mathbb{R}^d$:

$$s^P_{ik} = \text{Transformer}(s_{ik}, P, P) \tag{2}$$

where s_{ik} and P denotes the query and key/value in Transformer. In this way, for the input report y_i, we can obtain K phenotype-based textual representations, $\{s^P_{i1}, s^P_{i2}, \dots, s^P_{iK}\}$.

(iii) We adopt the phenotype-based textual representations s^P_{ik} to extract phenotype-based visual representations $v^P_{ik} \in \mathbb{R}^d$:

$$v^P_{ik} = \text{Transformer}(s^P_{ik}, V, V) \tag{3}$$

where V denotes the extracted patch features of the input medical image (Chen et al., 2020c). Through above equation, we can obtain K phenotype-based visual representations $\{v^P_{i1}, v^P_{i2}, \dots, v^P_{iK}\}$.

(iv) Now, for each pair of image and report (x_i, y_i), we can obtain K pairs of phenotype-based visual representations and textual representations, i.e., $\{(v^P_{i1}, s^P_{i1}), (v^P_{i2}, s^P_{i2}), \dots, (v^P_{iK}, s^P_{iK})\}$. The increased training samples indicate that our proposal can scale up contrasting learning.

During training, given a batch of N^* training samples, the phenotype-based contrastive learning (PCL) is defined as follows:

$$\begin{aligned}
\mathcal{L}^{v\rightarrow s} &= -\frac{1}{N^*}\sum_{k=1}^{N^*}\log\frac{\exp(\mathbf{sim}(v^P_k, s^P_k)/\tau)}{\sum_{j=1}^{N^*}\exp(\mathbf{sim}(v^P_k, s^P_j)/\tau)}\\
\mathcal{L}^{s\rightarrow v} &= -\frac{1}{N^*}\sum_{k=1}^{N^*}\log\frac{\exp(\mathbf{sim}(s^P_k, v^P_k)/\tau)}{\sum_{j=1}^{N^*}\exp(\mathbf{sim}(s^P_k, v^P_j)/\tau)}\\
\mathcal{L}_{\text{PCL}} &= \frac{1}{2}\left(\mathcal{L}^{v\rightarrow s} + \mathcal{L}^{s\rightarrow v}\right),
\end{aligned} \tag{4}$$

Through the above equation, we can enable PhenotypeCLIP to perform phenotype-based contrastive learning, learning fine-grained (phenotype-based) representations to achieve improved performance.

3.3 Report Generation

Medical report generation aims to automatically generate a medical report y given the input medical image x. Therefore, to perform the medical report generation, we follow previous works (Li et al., 2018; Jing et al., 2018; Chen et al., 2021, 2020c; Liu et al., 2021b,a) to adopt the encoder-decoder-based framework. We adopt the PhenotypeCLIP as the image encoder to extract the fine-grained visual representations, and the memory-driven Transformer (Chen et al., 2020c) as the decoder to generate accurate medical reports.

Given the ground-truth report $y^* = \{y^*_1, y^*_2, \dots, y^*_T\}$ for the input image x, the medical report generation model can be trained using the cross-entropy (XE) loss:

$$L_{\text{XE}}(\theta) = -\sum_{t=1}^{T}\log\left(p_\theta\left(y_t \mid y_{1:t-1}; x; \theta\right)\right) \tag{5}$$

4 Experiments

We first describe two benchmark datasets, the metrics, and the settings used for evaluation. Then, we present the main results and analysis of our proposal on the two datasets.

[1] Please refer to Vaswani et al. (2017) for details.

Methods	MIMIC-CXR				IU X-ray			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
AdaAtt (Lu et al., 2017)	0.088	0.118	0.266	0.084	0.068	-	0.308	0.295
TOPDOWN (Anderson et al., 2018)	0.074	-	0.250	0.073	-	-	-	-
Transformer (Vaswani et al., 2017)	0.090	0.125	0.265	-	0.135	0.164	0.342	-
R2Gen (Chen et al., 2020c)	0.103	0.142	0.277	-	0.165	0.187	0.371	-
PPKED (Liu et al., 2021b)	0.106	0.149	0.284	0.237	0.168	0.190	0.376	0.351
DeltaNet (Wu et al., 2022)	0.114	-	0.277	**0.281**	0.184	-	0.379	**0.802**
XProNet (Wang et al., 2022a)	0.105	0.138	0.279	-	0.199	0.220	0.411	0.359
MedCLIP* (Wang et al., 2022b)	0.109	0.146	0.283	0.255	0.178	0.204	0.382	0.347
ConVIRT* (Wang et al., 2022a)	0.101	0.142	0.275	0.249	0.175	0.196	0.380	0.341
PhenotypeCLIP	**0.119**	**0.158**	**0.286**	0.259	**0.205**	**0.223**	**0.414**	0.370

Table 1: Results of our approach, existing medical report generation models, and two conventional contrastive learning models on the two benchmark datasets. * denotes the re-implementations of existing contrastive learning methods for medical report generation.

4.1 Datasets, Metrics, and Settings

Datasets We performed the evaluation using the commonly used MIMIC-CXR dataset (Johnson et al., 2019) and IU X-ray dataset (Demner-Fushman et al., 2016). The MIMIC-CXR dataset comprises 377,110 chest X-ray images and 227,835 related radiology reports, whereas the IU X-ray dataset consists of 7,470 chest X-ray images and 3,955 reports. We follow Chen et al. (2020c) to pre-process the datasets: we adopt the official split to split the MIMIC-CXR dataset. The IU X-ray dataset is randomly split into training, validation and test sets by 7:1:2 of the entire dataset.

Metrics To report the performance of models for medical report generation, we use common evaluation metrics, i.e., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015).

Settings Following Zhang et al. (2020b), we use the ResNet-50 (He et al., 2016), which is pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on CheXpert (Irvin et al., 2019), as the image encoder to acquire the visual path features of input medical images, and the BERT (Devlin et al., 2019), which is initialized with the ClinicalBERT weights (Alsentzer et al., 2019), as the text encoder to acquire the textual features of input reports and sentences. We adopt the three-layer Transformer (Vaswani et al., 2017) to extract the phenotype-based representations. The number of attention heads and the hidden size d are set to 8 and 512, respectively. We adopt the memory-driven Transformer (Chen et al., 2020c) as the text decoder to generate final medical reports. For the phenotypes-based contrastive learning, inspired by the success of memory vectors (Chen et al., 2020c, 2021; Liu et al., 2021c), we adopt the randomly initialized soft vectors to implement the phenotypes. The number of phenotypes M is set to 100. The hyper-parameter τ is set to 0.1. During PhenotypeCLIP training, we adopt the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 1e-4, weight decay of 1e-6, and batch size of 32. During report generation model training, the learning rate and batch size are respectively set to 1e-4 and 16. During inference, we apply a beam search of size 3.

4.2 Results

The results of the proposal on the MIMIC-CXR and IU X-ray datasets are reported in Table 1. For comparison, we report the results of existing strong medical report generation models, Additionally, we re-implement two contrastive learning methods, i.e., MedCLIP (Wang et al., 2022b) and ConVIRT (Wang et al., 2022a). As we can see, the two conventional contrastive learning models (MedCLIP and ConVIRT) achieved competitive results with specific medical report generation models, suggesting that contrastive learning can provide a solid basis for report generation by bridging the gap between visual and textual modalities.

The proposed PhenotypeCLIP method demonstrated encouraging performances on the two benchmark datasets. On the MIMIC-CXR/IU X-ray datasets, PhenotypeCLIP achieved the highest BLEU-4 (0.119/0.205), METEOR (0.158/0.223), and ROUGE-L (0.286/0.414) scores, surpassing all other methods. In particular, our PhenotypeCLIP consistently outperformed all previous conventional contrastive learning methods across all evaluation metrics. For example, the PhenotypeCLIP outperforms previous contrastive learning models

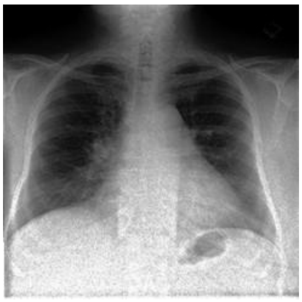
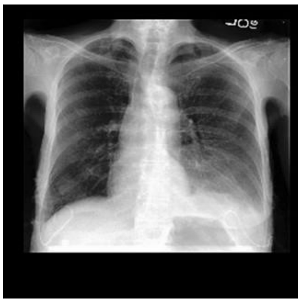
	**Ground Truth:** Both lungs are well expanded and clear. There are no lung opacities concerning for pneumonia or pulmonary edema. **Heart size is mildly enlarged** and stable since. Mediastinal and hilar contours are unchanged. There is no pleural effusion or pneumothorax.	**ConVIRT:** The cardiac, mediastinal and hilar contours appear stable. The lungs appear clear. No focal consolidation, pleural effusion or pneumothorax is seen. The heart is normal in size. Normal cardiomediastinal contours.	**PhenotypeCLIP:** The lungs are clear without evidence of focal consolidations concerning for pneumonia. **The heart size is enlarged.** No lung nodules or masses**.** There is no pleural effusion, pneumothorax, or evidence of pulmonary edema.
	**Ground Truth:** Bilateral pleural catheters remain in place, with persistent **pneumothoraces**, moderate left apical lateral pneumothorax on the left and small on the right. The left **pneumothorax** is unchanged, but right **pneumothorax** has minimally increased. Heart size remains normal. Persistent **left basilar atelectasis** and adjacent small **left pleural effusion**.	**ConVIRT:** The lungs are clear. The heart size is normal. The pulmonary vascularity is normal and the lungs are clear. There is no large pleural effusion or pneumothorax. There are no acute osseous abnormalities. There is no focal consolidation.	**PhenotypeCLIP:** Lung volumes remain low. The appearance of the cardiac silhouette is unchanged. There is a moderate **pneumothorax** evidenced in the right. Small bilateral **pleural effusions** are present. The mediastinal contours are within normal limits.

Figure 1: For a better understanding of our approach, we demonstrate the medical reports generated by a strong baseline model ConVIRT (Zhang et al., 2020b) and the proposed PhenotypeCLIP. We adopt the Bold text and Underlined text to denote the Correct results and Unfavorable results, respectively.

by 1.8% and 3.0% BLEU-4 score on MIMIC-CXR and IU X-ray datasets, respectively. It proves the effectiveness of our approach in learning fine-grained representations to boost the downstream medical report generation task.

4.3 Analysis[2]

In Figure 1, a qualitative analysis is conducted to give a better understanding of our PhenotypeCLIP. Specifically, we show two medical reports generated by a strong contrastive learning model ConVIRT (Zhang et al., 2020b) and our method. It is clear that our method can generate better reports than the ConVIRT on the two input images. For example, in the first example, given the ground truth {Heart size is mildly enlarged}, ConVIRT gives a wrong description, while PhenotypeCLIP accurately describes the abnormality, i.e., {The heart size is enlarged.}. In the second example, ConVIRT is unable to capture any anomalies and incorrectly describes the input image in a normal case. Fortunately, although our proposal can not capture the "basilar atelectasis", it correctly describes the "pneumothorax" and "pleural effusions". The encouraging results demonstrate that our phenotype-based contrastive learning can efficiently improve medical report generation by learning accurate and fine-grained phenotype-based representations, which are helpful in capturing and describing the abnormalities. It further proves the effectiveness of our proposed approach.

5 Conclusion

In this work, we propose a phenotype-based (fine-grained) contrastive learning framework, PhenotypeCLIP, for medical imaging report generation. The proposed PhenotypeCLIP can efficiently bridge the gap between visual and textual modalities to provide a solid basis for generating accurate medical reports. In the implementation, PhenotypeCLIP learns fine-grained representations, i.e., phenotype-based representations, by performing contrastive learning on the image and each sentence within the report. Experiments on two widely-used datasets MIMIC-CXR and IU X-ray datasets prove our arguments and show that PhenotypeCLIP achieves competitive results with previous state-of-the-art methods, especially exhibiting a remarkable improvement over conventional contrastive learning techniques across all evaluation metrics.

In future research, it is interesting to explore extracting the topics or keywords, e.g., abnormalities (Jing et al., 2018), from the reports to build the set of phenotypes.

[2] Please refer to our supplementary material for more analysis, e.g., the sensitivity of hyper-parameters and examples.

Acknowledgments

We deeply thank all the anonymous reviewers and chairs for their helpful comments and suggestions that substantially improved this paper.

Limitations

The study heavily relies on the quality and amount of the labeled medical report generation datasets for training. It might be possible that in some cases, a large-scale labeled dataset might not be available, leading to unsatisfactory results. Besides, the current phenotype set might not cover all the phenotypes or abnormalities, e.g., rare diseases. Therefore, performance could be further improved by expanding and refining the phenotypes. At last, the second example in Figure 1 shows that our approach cannot well capture the details of abnormalities, such as the position ('left' and 'bilateral'). Introducing a position prediction module or knowledge graph (Yang et al., 2022) could be helpful in addressing this problem.

Ethics Considerations

The two benchmark datasets we used are publicly available, so no protected health information is disclosed. Any inaccuracies, including misdiagnosis or missed abnormalities, in the generated reports can lead to incorrect clinical outcomes. It's essential to control the use of model-generated reports. It's crucial to ensure that medical professionals review and validate the generated reports in clinical practice. At last, similar to any existing deep learning models, PhenotypeCLIP is vulnerable to inherent biases in training data. Ensuring fairness and avoiding potential bias is critical.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

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*.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, pages 1–21. Springer.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *ACL/IJCNLP*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020c. Generating radiology reports via memory-driven transformer. In *EMNLP*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc.*, 23(2):304–310.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: pre-trained prompt tuning for few-shot learning. In *ACL*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *International Conference on Computer Vision*, pages 3922–3931.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, 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*.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*.

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, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *EMNLP*.

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*.

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*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*.

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*.

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*.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *MICCAI*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020a. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020b. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. 2023. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*.