Few-shot Structured Radiology Report Generation Using Natural Language Prompts

Matthias Keicher
Technische Universität München
28 PUBLICATIONS **202** CITATIONS
SEE PROFILE

Kamilia Zaipova
Technische Universität München
4 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Tobias Czempiel
Technische Universität München
31 PUBLICATIONS **265** CITATIONS
SEE PROFILE

Kristina Mach
Technische Universität München
7 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Few-shot Structured Radiology Report Generation Using Natural Language Prompts

Matthias Keicher[1], Kamilia Mullakaeva[1], Tobias Czempiel[1], Kristina Mach[1], Ashkan Khakzar[1], Nassir Navab[1,2]

[1] Computer Aided Medical Procedures, Technische Universität München, Germany
[2] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. Chest radiograph reporting is time-consuming, and numerous solutions to automate this process have been proposed. Due to the complexity of medical information, the variety of writing styles, and free text being prone to typos and inconsistencies, the efficacy quantifying the clinical accuracy of free-text reports using natural language processing measures is challenging. On the other hand, structured reports ensure consistency and can more easily be used as a quality assurance tool. To accomplish this, we present a strategy for predicting clinical observations and their anatomical location that is easily extensible to other structured findings. First, we train a contrastive language-image model using related chest radiographs and free-text radiological reports. Then, we create textual prompts for each structured finding and optimize a classifier for predicting clinical findings and their associations within the medical image. The results indicate that even when only a few image-level annotations are used for training, the method can localize pathologies in chest radiographs and generate structured reports.

Keywords: Structured report generation · Contrastive training · Few-shot learning

1 Introduction

Structured reports are increasingly used in the radiology field and are endorsed by large radiological societies such as RSNA and ESR [20]. By standardizing content and terminology, structured reports simplify and streamline the communication and make the reports machine-readable, allowing for better quality assurance, data analysis, and indexing [10]. Most importantly, they reduce the workload for the medical staff by partially automating this time-consuming task. With the rise of powerful Natural Language Processing (NLP) models like transformers [27] and the availability of large chest radiograph datasets, many methods have been proposed for the generation of free-text radiology reports. However, these generated reports suffer from the same drawbacks as human-written free-text reports: the terminology is not standardized, and therefore it is difficult to evaluate whether a report is clinically accurate or merely performs well on common NLP metrics which do not measure the expressiveness of a report in all its clinical details [24]. To this end, we propose a structured report generation method that requires less annotated data by leveraging pre-training of a contrastive language-image model on chest radiograph and report pairs. Our contributions are:

- a novel few-shot model for structured report generation of chest radiographs based on flexible templates that can easily be adapted with textual prompts.
- a data-efficient method for contrastive image-language pre-training suitable for chest radiographs and reports.
- an extensive evaluation on the model in a few-shot setting and comparison with models on two tasks: pathology localization and severity grading of cardiomegaly.

Related work To create a medically expressive report Liu et al. [19] and Yang et al. [31] make us of posterior and prior clinical knowledge. In Endo et al. [8] the use of a fine-tuned contrastive language-image pretraining (CLIP) model [26] for radiology report generation was proposed for report retrieval. Yan et al. suggested a weakly supervised contrastive learning pipeline where labels were created using a k-mean report cluster [30]. Further, the hierarchical structure of the pathologies were modeled [23,6]. Another set of works make use of Graph convolutional networks (GCN) to exploit label dependencies [32,5]. Zhang et al. suggested a report generation informed by a previously constructed disease knowledge graph to jointly learn features and model relationships [32].

However, only a few works focus their research on structured report generation. Pino et al. [25] proposed a model for structured reporting generation where abnormalities are first classified, and the correct report templates are automatically selected. To improve the structured reporting results, additional object detectors were proposed as an additional step [3]. For localization, Jaiswal et al. [13] proposed to highlight abnormal regions with class activation maps.

Under low data regimes and long-tailed distributions few-shot learning approaches have been developed and applied to the task of radiograph examination [21,22,14]. With RareGen, Jia et al. [14], proposed a report generation model for rare diseases, utilizing few-shot learning. To improve pulmonary edema classification [4] make us of pretraining with free-text reports and radiographs. Similarily [18] leverage multimodal representation by maximizing local mutual information.

2 Method

Our proposed Few-shot Structured Report Generation (FSRG) method leverages contrastive image-language pretraining (CLIP) by projecting all sentences of a structured radiology report in a joint embedding space, predicting the combination of sentences that are most similar to the input image embedding. Inspired by the success in detection-free human-object interaction recognition (DEFR) [15], our method consists of a text and image encoder with images I as input and multiple class labels $y = \{y_1, y_2, \dots, y_C\}$ as output where C is the number sentences, and $y_i \in \{1, -1\}$, indicating positive and negative classes, respectively.

The text encoder transforms each possible sentence r_i of the structured report to a language embedding. These embeddings are used to initialize the weights of our structured report classifier formed by a fully connected layer without bias with the number of sentences as an output size. To classify the chest radiographs, they are first encoded with an image encoder and then projected with a fully connected layer without bias to the language embedding space. By normalizing the weights and input of the classifier, the classifier calculates the cosine similarity s_i between the image and text embedding creating the output logits for each sentence.

Most textual prompts have common information with other sentences from similar parts of the template. For example, *lung opacity in the left lung* and *lung opacity in the upper left lung* have a almost identical language embedding, naturally modeling label dependencies and providing a initialization on which one can build on. Finally, all similarities are processed to exclusive options using softmax.

2.1 Log-Sum-Exp Sign loss

We adopt a Log-Sum-Exp Sign (LSES) loss function inspired by [15], to account for the fact that the majority of results in a structured report are non-exclusive. γ is a hyper-parameter scaling the output range of the similarity s_i. The overall loss is defined as:

$$\mathcal{L} = \log\left(1 + \sum_{i=1}^{C} e^{-y_i \gamma s_i}\right) \tag{1}$$

The loss automatically puts a higher weight on misclassified classes, improving the learning of the long-tail distribution for chest radiograph datasets.

2.2 Contrastive language-image pretraining

There are multiple choices for the combination of text and image encoders. One option is to use pretrained domain-specific models, trained independently.

The large-scale image-language model CLIP [26] trained on natural image-text pairs showed the best performance in human-object interactions [15]. The advantage of this approach is that text and image embeddings should already be aligned similarly since their contrastive representations have been trained with cosine similarity. No models are available yet that have been trained on a similar scale for chest radiograph and report pairings. We, therefore, investigate several setups of fine-tuning CLIP to chest radiographs, and training a similar model from scratch.

3 Experimental setup

Dataset We use the MIMIC-CXR-JPG v2.0.0 [16] dataset, which is derived from the MIMIC-CXR dataset consisting of 377,110 chest radiographs associated with 227,827 imaging studies and free-text reports [17,9]. The labels for

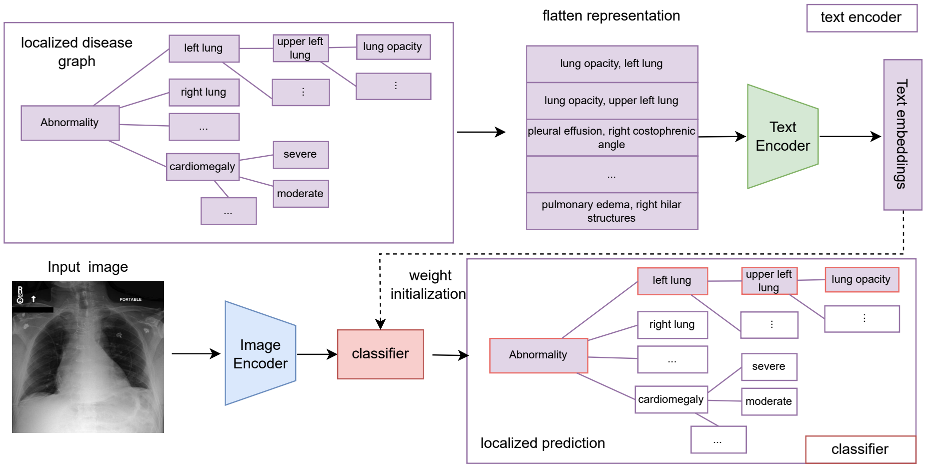

Fig. 1. Few-shot Structured Report Generation (FSRG): weights of the classifier are initialized text embeddings from the template elements defined by the textual prompts. The classification module calculates the cosine similarity between the image embedding and class weights. Finally, the similarities of all classes are fed to the template tree modeling label dependencies. With the filled template, a structured report can be generated.

structured reports are extracted from the medical scene graph dataset Chest ImaGenome [29,28] consisting of 242072 anatomy-centered scene graph for the MIMIC-CXR image data. It provides 1256 combinations of relation annotations between 29 anatomical locations and their attributes. We use the data split provided by ImaGenome with Posterior Anterior (PA) and Anterior Posterior (AP) radiographs resulting in 166512 training images, 23952 validation images and 47389 test images after preprocessing. The images were processed with MONAI 0.8.0 and the dataloader is implemented with ffcv 0.0.2. All images are resized to 224x224, padded if needed and scaled to the range [-1,1]. For training, the following image augmentations were applied: random crop with at least 75% image size, random rotation up to ±15° and a color jitter of 10% brightness as well as 20% contrast and saturation. The reports were augmented by randomly sampling a sentence containing a finding in the ImaGenome scene graph. For healthy patients with no findings, a random sentence from the full report was randomly sampled.

Implementation details All networks were trained with PyTorch 1.10 and PyTorch lightning 1.5.10 in native mixed precision. The transformer-based models and tokenizers are implemented using the huggingface library (transformers 4.16.2). The CLIP models were fine-tuned on 8 NVIDIA A40 for 300 epochs with a batch size of 128 using an AdamW optimizer with a learning rate of 5e-6, no weight decay on normalization layer and bias and 0.1 weight decay on all other parameters. The learning rate was decayed with a cosine annealing schedule and 1 epoch linear warmup. All other models were trained on a single NVIDIA A40. The DenseNet and Vision Transformer classifiers were trained for 25 epochs with the same hyperparameters as in [1]: Adam optimizer, a learning rate of 1e-4 and unweighted binary cross-entropy loss. The FSRG models were fine-tuned for 10 epochs with a learning rate of 1e-4, AdamW optimizer with no weight decay, and learning rate scheduler with cosine annealing decay and 1 epoch linear warmup. The model performing best on the validation set was used for testing. After a hyperparameter search on the localized pathology task with 50, 100 and 150 the γ parameter of the LSES loss was set to 50 for all experiments. For the few-shot learning, an epoch is defined as seeing 128 images per class. For the setup of sampling all classes, the same amount of images was chosen per class to ensure comparability while having access to the entire dataset. A batch size of 256 is used for the classification baselines and FSRG finetuning. All experiments were repeated 5 times with different seeds and the average results are reported.

3.1 Localized pathology detection

To evaluate our method we test it on two structured reporting tasks. The first task is localizing pathologies in chest radiographs using the MIMIC-CXR dataset. For this we use the labels provided by ImaGenome [28], that have also been used in [1]. The 9 attributes are: *Lung Opacity*, *Pleural Effusion*, *Atelectasis Enlarged Cardiac Silhouette*, *Pulmonary Edema/Hazy Opacity*, *Pneumothorax*, *Consolidation*, *Fluid Overload/Heart Failure* and *Pneumonia*. The dataset contains 19 anatomical locations including different regions of the lung, hilar structures, costophrenic angle and mediastinum as well as the cardiac silhouette and trachea. For each patient we extract the triplet of *attribute* located in the *anatomical site* from the medical scene graph provided. For all patients this resulted in 98 (of 162 possible) unique combinations of attribute and location. By joining the *attribute* and *location* with *"in the"* we create the template sentences used as an initialization of the classifier, for example *"consolidation in the left lung"*. Following [1] we calculate the area under the receiver operating characteristic curve (AUC) for all possible locations of each pathology and average them to one location-sensitive AUC per pathology. For fair comparison with the image encoder used in CLIP, the DenseNet121 used as a baseline for detection-free localization has been pretrained on the global pathology classification.

3.2 Severity prediction of cardiomegaly

The second task evaluates the model to predict the severity of cardiomegaly following the textual prompts defined in the TLAP endorsed structured report template *Chest Xray - 2 Views* created by Penn Medicine at the University of Pennsylvania[3]. Using the keywords we have extracted from reports 6 possible severity states of cardiomegaly. Prompts distribution can be seen in Table 1. We

[3] https://radreport.org/home/144/2011-10-21%2000:00:00

calculate probabilities using softmax and train the DenseNet121 baseline with a cross entropy loss. The methods are compared using the AUC.

Severity	Intialization prompt	Training	Validation	Testing
Normal	The heart is normal in size.	3140	478	943
Top Normal	The heart is top normal in size.	635	72	160
Mild	There is mild cardiomegaly.	6084	809	1816
Moderate	There is moderate cardiomegaly.	8696	1164	2619
Severe	There is severe cardiomegaly.	2231	335	676
Marked	There is marked cardiomegaly.	246	36	85
		21032	2894	6299

Table 1. The prompts were used for initializing the classifier in the second task of predicting the severity of cardiomegaly. Following the *Chest Xray - 2 Views* template from RadReport.

4 Results and discussion

4.1 Contrastive language-image pretraining

The original CLIP model is trained on pairs of natural images and captions and, therefore, not directly applicable for downstream tasks on radiology images. Endo et al. [8] fine-tuned CLIP on radiology images for report generation. We first adopted this strategy and fine-tuned a CLIP model with a ViT-16/B backbone [7] using the original CLIP tokenizer and text encoder. When reaching a training accuracy of above 20%, the model started overfitting on the training set after about 100 epochs, with the validation accuracy continuously decreasing from its maximum of 7%. Since the tokenizer of CLIP is built from a large corpus of words that are primarily not related to radiology reports, we replace the CLIP tokenizer with the byte-level Byte-Pair Encoding tokenizer used in RATCHET [11] and SciBERT(SB) [2] that have been trained on MIMIC-CXR reports and scientific text, respectively. This resulted in a slightly worse validation and training accuracy, indicating that the large-scale pretraining of CLIP is more critical than a domain-specific tokenizer. To challenge this idea, we replaced the text and image encoder with domain-specific encoders DenseNet121 (DN121) [12] trained on the classification of non-localized pathologies in the ImaGenome dataset (see section 3.1) and SciBERT pre-trained on a large corpus of scientific text. Surprisingly, this setup showed the best validation accuracy of 10% but a lower maximum training accuracy of 11%, indicating better generalizability. For all following setups, we used this model and compared it with the best performing, fine-tuned CLIP model with a CLIP tokenizer.

Method	Lung Opac.	Pleural Eff.	Atelectasis	Enl. Card. S.	Pulm. Edema	Pneumothor.	Consolidation	Heart Failure	Pneumonia	**Avg. AUC**
Global view with no localization										
DenseNet169 [1]	0.91	0.94	0.86	0.92	0.92	0.93	0.86	0.87	0.84	0.89
DenseNet169	0.87	0.90	0.79	0.86	0.85	0.83	0.75	0.77	0.75	0.82
DenseNet121	0.88	0.91	0.81	0.87	0.87	0.87	0.79	0.80	0.77	0.84
ViT-B16	0.88	0.91	0.80	0.87	0.86	0.85	0.77	0.78	0.76	0.83
Object detection backbone with high resolution crops										
FasterR-CNN [1]	0.84	0.89	0.77	0.85	0.87	0.77	0.75	0.81	0.71	0.80
AnaXNet [1]	0.88	0.96	0.92	0.99	0.95	0.80	0.89	0.98	0.97	0.93
Detector free localization on global view (224 × 224)										
DenseNet121 1-shot	0.70	0.76	0.64	0.77	0.70	0.60	0.66	0.62	0.58	0.67
DenseNet121 5-shot	0.72	0.78	0.66	0.78	0.73	0.64	0.67	0.64	0.62	0.69
DenseNet121 full	0.83	0.89	0.79	0.87	0.84	0.89	0.83	0.81	0.82	0.84
FSRG 1-shot	0.72	0.83	0.69	0.82	0.77	0.72	0.74	0.73	0.67	0.74
FSRG 5-shot	0.75	0.84	0.71	0.82	0.79	0.78	0.76	0.73	0.71	0.77
FSRG full	0.82	0.89	0.78	0.87	0.84	0.90	0.83	0.80	0.81	0.84

Table 2. Evaluation of our approach against baselines for the localized detection of pathologies in chest radiographs. Except for the global view baselines, the AUC scores for each pathology are averaged over all anatomical locations. The proposed method is marked bold with 1 and 2 images per class used for training. The full training indicates the use of the full dataset as in the global view and object detection methods. The scores of the cited models have been adopted.

Method	Backbone	Pretraining	1-shot	5-shot	10-shot	100-shot	sampled	full
Ablation on localized pathologies								
MLP	DN121	pathologies	0.67	0.69	0.71	0.76	0.77	0.84
FSRG	ViT-B/16-CLIP	CLIP	0.66	0.70	0.73	0.75	0.77	-
FSRG	DN121+SB	random init.	0.67	0.72	0.75	0.79	0.81	-
FSRG	**DN121+SB**	**CLIP**	0.74	0.77	0.78	0.80	0.81	0.84
Grading task: Cardiomegaly severity prediction								
MLP	DenseNet121	pathologies	0.59	0.65	0.68	0.75	0.79	0.73
FSRG	**DN121+SB**	**CLIP**	0.65	0.72	0.74	0.77	0.78	0.82

Table 3. Comparison with naïve transfer learning using an MLP and ablation of the model with different backbones and initializations. Random initialization refers to the classifier weights that are initialized with textual prompts in FSRG. The grading task shows the method's performance for exclusive choices in a template with severity grading of cardiomegaly (AUC score). Highlighted in bold is the proposed approach.

4.2 Few-shot structured report generation

Localized pathology detection In the second set of experiments, we investigate the effectiveness of FSRG in predicting structured labels with a limited amount of training samples per class. We use the pre-trained CLIP model and fine-tune FSRG to the task-specific textual prompts. Table 2 shows the results for localized pathology detection, where FSRG is compared with global classification baselines and methods based on object detection that make use of the full training data. The image encoders evaluated for global pathology detection perform similarly with an average AUC of 83%. The DenseNet169 performance reported in [1] could not be reached, which we assume might be due to our limited image resolution of 224×224. Neither FSRG nor the naïve transfer learning baseline reach the AUC of AnaXNet, which is built upon an object detector and uses features extracted from high-resolution crops. However, FSRG has a 0.07 higher AUC for 1-shot learning and 0.08 increase for 5-shot learning in comparison with transfer learning with DenseNet121. Learning the generation of structured reports with limited data is clinically useful since high-quality data annotations - particularly on a voxel level - are time-consuming, and the templates of FSRG could easily be adapted to hospital-specific reporting or changing guidelines.

Ablation and severity grading Table 3. shows that the CLIP model with ViT-B/16 backbone that overfitted during training has a similar performance in few-shot learning as the DenseNet121 baseline and an FSRG model initialized with random weights instead of textual prompts projected to the class embeddings. This highlights the importance of the initialization by the pretrained language-image mode with meaningful clustering of semantically related concepts. With an increasing number of samples seen per class, the gap between FSRG and the other models decreases. In the task of grading cardiomegaly, a similar pattern can be seen with an increase of AUC by 0.06 in the 1-shot case and 0.07 in 5-shot learning over the naïve transfer learning baseline.

5 Conclusion

This paper proposes a method for generating structured reports from chest radiographs in a low data regime. We compared the performance on different few-shot learning settings, limiting the number of image-level annotations and demonstrated an increased performance over the transfer learning baseline. In this work we show a step towards pathology localization without the use of additional object detection backbones or pixel-level annotations. Next, we want to investigate the influence of improving the contrastive pretraining and modeling of the entire medical scene graph as a flexible structured template.

Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under project DIVA (FKZ 13GW0469C). Ashkan Khakzar was partially supported by the Munich Center for Machine Learning (MCML) with funding from the BMBF under the project 01IS18036B. Kristina Mach was partially supported by the Linde & Munich Data Science Institute, Technical University of Munich PhD Fellowship.

References

1. Agu, N.N., Wu, J.T., Chao, H., Lourentzou, I., Sharma, A., Moradi, M., Yan, P., Hendler, J.: Anaxnet: Anatomy aware multi-label finding classification in chest x-ray. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 804–813. Springer (2021)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
3. Bhalodia, R., Hatamizadeh, A., Tam, L., Xu, Z., Wang, X., Turkbey, E., Xu, D.: Improving pneumonia localization via cross-attention on medical images and reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 571–581. Springer (2021)
4. Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–539. Springer (2020)
5. Chen, B., Li, J., Lu, G., Yu, H., Zhang, D.: Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. IEEE Journal of Biomedical and Health Informatics **24**(8), 2292–2302 (2020). https://doi.org/10.1109/JBHI.2020.2967084
6. Chen, H., Miao, S., Xu, D., Hager, G.D., Harrison, A.P.: Deep hierarchical multi-label classification of chest x-ray images (08–10 Jul 2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
8. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Roy, S., Pfohl, S., Rocheteau, E., Tadesse, G.A., Oala, L., Falck, F., Zhou, Y., Shen, L., Zamzmi, G., Mugambi, P., Zirikly, A., McDermott, M.B.A., Alsentzer, E. (eds.) Proceedings of Machine Learning for Health. Proceedings of Machine Learning Research, vol. 158, pp. 209–219. PMLR (04 Dec 2021)
9. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)
10. Hong, Y., Kahn, C.E.: Content analysis of reporting templates and free-text radiology reports. Journal of digital imaging **26**(5), 843–849 (2013)

11. Hou, B., Kaissis, G., Summers, R.M., Kainz, B.: Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 293–303. Springer (2021)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
13. Jaiswal, A., Li, T., Zander, C., Han, Y., Rousseau, J.F., Peng, Y., Ding, Y.: Scalp-supervised contrastive learning for cardiopulmonary disease classification and localization in chest x-rays using patient metadata. In: 2021 IEEE International Conference on Data Mining (ICDM). pp. 1132–1137. IEEE (2021)
14. Jia, X., Xiong, Y., Zhang, J., Zhang, Y., Zhu, Y.: Few-shot radiology report generation for rare diseases. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 601–608. IEEE (2020)
15. Jin, Y., Chen, Y., Wang, L., Wang, J., Yu, P., Liang, L., Hwang, J.N., Liu, Z.: Decoupling object detection from human-object interaction recognition. arXiv preprint arXiv:2112.06392 (2021)
16. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). PhysioNet (2019)
17. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
18. Liao, R., Moyer, D., Cha, M., Quigley, K., Berkowitz, S., Horng, S., Golland, P., Wells, W.M.: Multimodal representation learning via maximization of local mutual information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–283. Springer (2021)
19. 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)
20. Nobel, J.M., van Geel, K., Robben, S.G.F.: Structured reporting in radiology: a systematic review to explore its potential. European Radiology (Oct 2021)
21. Paul, A., Shen, T.C., Peng, Y., Lu, Z., Summers, R.M.: Learning few-shot chest x-ray diagnosis using images from the published scientific literature. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 344–348 (2021)
22. Paul, A., Tang, Y.X., Shen, T., Summers, R.: Discriminative ensemble learning for few-shot chest x-ray diagnosis. Medical Image Analysis **68**, 101911 (02 2021)
23. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
24. Pino, P., Parra, D., Besa, C., Lagos, C.: Clinically correct report generation from chest x-rays using templates. In: International Workshop on Machine Learning in Medical Imaging. pp. 654–663. Springer (2021)
25. Pino, P., Parra, D., Besa, C., Lagos, C.: Clinically correct report generation from chest x-rays using templates. In: MLMI@MICCAI (2021)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Syeda-Mahmood, T., Moradi, M.: Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316 (2021)
29. Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Syeda-Mahmood, T., Moradi, M.: Chest imagenome dataset (version 1.0.0). PhysioNet (2021)
30. Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.N.: Weakly supervised contrastive learning for chest x-ray report generation. arXiv preprint arXiv:2109.12242 (2021)
31. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Radiology report generation with general and specific knowledge. arXiv preprint arXiv:2112.15009 (2021)
32. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)