Improving Medical Knowledge in the Automated Chest Radiograph Report

Ethan Schonfeld
Department of Biology
Stanford University
eschon22@stanford.edu

Edward Vendrow
Department of Computer Science
Stanford University
evendrow@stanford.edu
Not Enrolled in CS230

Abstract

The clinical writing of unstructured reports from chest radiograph imaging is error prone, due to the lack of its standardization and repeated report writing daily, which can prove fatal. A system that generates reports can assist clinicians to reduce errors. Such a system could further be used as a training tool for medical education as well as used in the global setting to promote medical accessibility in low resource areas. Current medical report generating efforts employ the BLEU metric which was shown to score better clinically meaningless, yet grammatical, random reports [1–2]. Further, state of the art methods for this task pretrain the visual extractor on Imagenet which has been shown to generalize poorly for medical domain applications [3]. We seek to study the benefit of pretraining on a chest radiograph specific trained visual extractor. We also combine both feature extractors to study how this extra input information to the generating model can improve the semantic medical accuracy of the resulting reports. We test each model by evaluating BLEU(1–4) metrics and F1 score performance on each of 14 possible labels as defined by CheXpert for chest radiographs. We find that while the chest radiograph feature extrator model and the double feature model result in lower BLEU scores, they perform better across specific F1 scores and total F1 score. We provide evidence to suggest that the choice of Imagenet, domain specific, or combined feature extractor is dependent specifically on which medical knowledge is most important for the application. This supports the further investigation of using a combined domain specific feature extractor with an Imagenet pretrained feature extractor for medical imaging captioning tasks.

1 Introduction and Related Work

In the medical domain, a task that appears in almost all specialities is the generation of reports from medical imaging. Whether this imaging is simple 2D chest radiographs or 3D time series of functional brain activity mappings, experienced clinicians generating many such reports daily are error prone. In a medical setting, such errors could prove fatal. Advances in deep learning based image captioning allow for the potential automation of such clinical tasks.

The captioning and report generation for chest radiographs is an emerging area of research with the release in 2019 of MIMIC CXR, a dataset composed of over 220,000 studies with chest radiographs and their associated clinical report [4]. State of the art models for such report generations have used transformers, transformer coupled with relational memory, LSTM with reinforcement learning, and retrieval techniques [5-7].

Current approaches use pretraining on Imagenet for the encoding of the chest radiograph as input to the generation model. However, recent work has demonstrated that Imagenet pretraining does not transfer well to medical domain tasks [3]. For high parameter models in the medical domain, Imagenet pretraining does not provide a large boost, and more parameter efficient models for chest radiograph specific tasks can be constructed without Imagenet transfer learning [3]. This gap in Imagenet feature learnings and medical domain features was shown to be significantly important in later task performance [8]. In recent work, using radiographs to learn from labelled manual annotations, thereby replacing Imagenet, resulted in outperforming Imagenet based state of the art models [8]. CheXpert is a large dataset of chest radiographs matched with labels for 14 medical conditions: Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, No Finding [9]. The large competition on CheXpert performance gives the opportunity to use a domain specific feature extractor for chest radiograph input for later NLP tasks.

While state of the art methods, using Imagenet, have resulted in strong performance on standard natural language generation metrics such as BLEU score, evaluation is significantly hampered by the domain agnosticism of such metrics. Novel research demonstrated that such metrics assign high scores to grammatically correct, yet, clinically irrelevant models [1, 2]. Therefore, we explore evaluation techniques to emphasize medical semantic performance and study how different feature extraction methods impact such medical semantic model performance.

The release of MIMIC-CXR dataset inspired multiple efforts for chest radiograph image captioning. The field existed prior to the 2019 release; however, it was limited to datasets using only a few thousands matched image to report examples. With the release of MIMIC-CXR containing over 200,000, more advanced models making use of novel transformer architectures catalyzed performance in the growing field. The state of the art model, as quantified by BLEU Scores, used a transformer architecture. It modifies the standard transformer to incorporate the concept of relational memory by allowing the model to pay attention to past cycles during a generating cycle [5]. The improved performance of this transformer based generating model inspired our approach to be transformer based, rather than past LSTM and RL models [6,7, 10].

2 Approach

All models follow the format as depicted in Figure 1, using a visual feature extractor CNN to provide a series of vector inputs representing positional features of the radiograph to the generating model.

Our baseline is a transformer with 6 encoder layers and 6 decoder layers that uses an Imagenet pretrained densenet121 CNN as a visual feature extractor. We refer to this model as (IMG:TF) for Imagenet pretrained Transformer. This was coded ourselves with significant modifications and revisions to starter code from CATR Image Captioning [11]. Visual feature extraction was coded ourselves with extracting layers and reshaping from PyTorch's densenet121 pretrained model. We use state of the art metrics for this task from recent work using transformer with relational memory, a component that records memory information across the generation process [5].

The next constructed model is identical to IMG:TF except for its visual feature extractor. Rather than using the Imagenet pretrained CNN to extract visual features (64x256 layer matrix) we extract the same dimension layer from a CheXpert pretrained model. This model, ranked 5th in the CheXpert leaderboard was trained to predict the presence of the 14 CheXpert labels from a chest radiograph input [12]. This was used as the feature extractor in our generating model, and the 64x256 feature vectors were used as input to the encoding block of the transformer, just as in IMG:TF. This model we refer to as (CHX:TF) for CheXpert pretrained Transformer.

Our last model was built to investigate the potential benefit of inputting information trained on Imagenet as well as information trained on CheXpert. By doing so object recognition as well as medical knowledge may benefit the report generation process. To accomplish this, we investigated how multimodal inputs are best incorporated in a transformer model. Each series of feature inputs is passed through its own encoder block composed of encoding layers (Figure 2). Recent work has explored whether these encoding layers should be concatenated or inputted "serially" each into its own attention block in decoding layer, one after the other. Best performance was achieved in serial models and thus is the design architecture that we use for the combined model [13]. This model we refer to as (IMG+CHX:TF) for Imagenet and CheXpert pretrained Transformer (Figure 2).

A beam search of size 5 was used for generating on evaluation sets for the IMG:TF and IMG+CHX:TF models, but not for the CHX:TF model as this required high compute and time resources and did not improve performance in initial experiments.

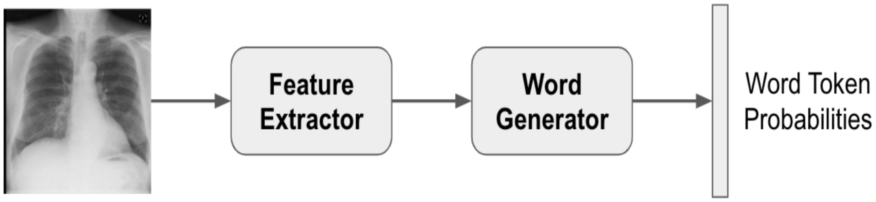


Figure 1: Base model for an automated chest radiograph report

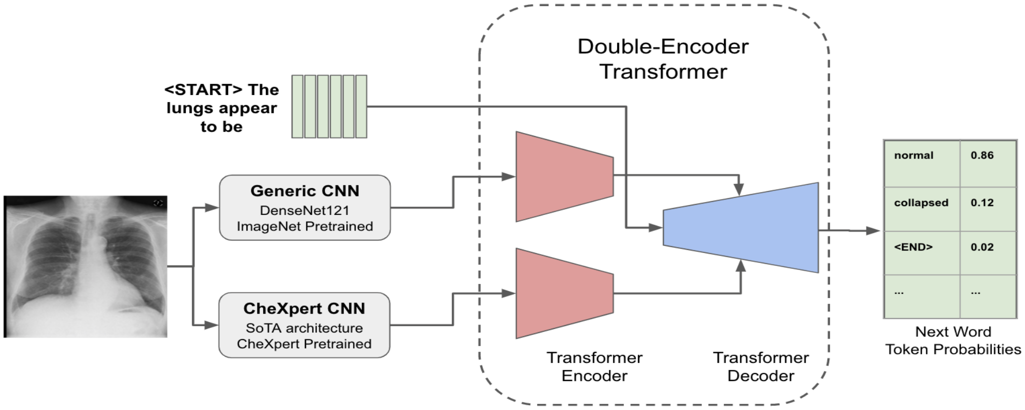


Figure 2: IMG+CHX:TF model for an automated chest radiograph report

3 Experiments

3.1 Data

We use the MIMIC–CXR dataset [4]. It consists of a 3 channel JPG 256x256 chest radiograph and associated clinical report by an expert radiologist. We use a training set composed of 152,173 radiographs and its distinct clinical report (only findings section); validation set of 1196, and testing set of 2347. The images are randomized to select a variety of frontal, lateral, PA, and AP views for the selected radiograph. Standard normalization using mean and standard deviation is applied across all images. GloVe word embeddings are used. GloVe embeddings are used in this work as word similarities were used to define embeddings for unseen words in the corpus and account for the most frequent medical typos. However, in later experiments we will use a frozen BioBERT model. The task is given a radiograph to generate its clinical report.

3.2 Evaluation method

To evaluate the generated reports we used the ground truth clinical reports and BLEU1,2,3,4 scores. As discussed, the BLEU metric has been shown to be domain agnostic and reward grammatically correct but clinically irrelavant models. As the field of radiology begins to move towards a structured report, a metric for correct labeling of common conditions and findings in the reports is required. To accomplish this we made us of the CheXbert Automatic Report Labeler [14]. CheXbert uses free text unstructured report as input and outputs the 14 CheXpert medical labels. Each label (condition/finding) is assigned a 1.0 (positive), 0.0 (negative), -1.0 (uncertain), or Blank (NaN). We extracted this set of 14 labels for the ground truth reports as well as generated reports for IMG:TF, CHX:TF, and IMG+CHX:TF models. For each label, and for each model, F1 score was calculated as well as the total F1 score for each model across all findings labels.

3.3 Experimental details

All three models were run for 10 epochs (until validation set loss plateaued) with batch size 64. IMG:TF and CHX:TF had one encoder block and one decoder block with 6 layers each. IMG+CHX:TF had two encoder blocks (Figure 2) with 6 layers each and one decoder block with 6 layers.

A learning rate scheduler was used to decrease learning rate as training continued and gradient clipping used. Adam optimizer was used. Each layer for both transformers multihead attention head count was 8. Learning rate for the model's backbone was 1e-05 and learning rate for all non–backbone components was 1e-04.

3.4 Results

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG:TF	0.233	0.140	0.0859	0.0544
CHX:TF	0.225	0.135	0.0836	0.0519
IMG+CHX:TF	0.209	0.125	0.0748	0.0494
Chen et al State Of The Art [5]	**0.353**	**0.218**	**0.145**	**0.103**

Figure 3: BLEU Metrics across models

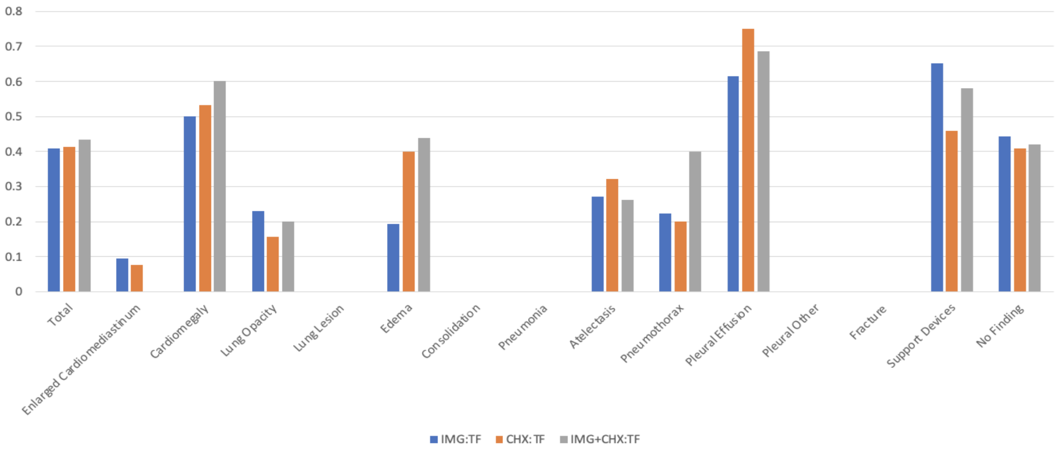


Figure 4: F1 Scores for Total and across the 14 CheXpert labels

The first thing to notice about the results is that the F1 score was not able to be calculated for some categories such as Enlarged Cardiomediastinum for IMG+CHX:TF or Consolidation for any of the models. The reason for this is that a random sample of 200 images from validation set, which is 1100 total images, were used for metric evaluation. These reports had a zero true positive value for prediction of these labels. CheXbert labeling assigns a score of -1.0 when it is uncertain. For our purposes we only considered a score of 1.0 to be a prediction. By including these -1.0 maybe predictions F1 scores for the remaining categories can be calculated but the degree of confidence we have in the quantitative values across all categories will be reduced. Next, we notice that by BLEU(1-4) metrics, IMG:TF is the superior model. However, by total F1, it is the worst model. This agrees with reports of the BLEU metric being domain agnostic in medicine and rewarding grammatic yet clinically irrelevant models [1-2]. It is likely that CheXpert visual extraction decreases the grammar quality but increases the knowledge level of the reports.

Interestingly, IMG:TF is superior by F1 metric for some specific label categories such as Support Device (such as pacemakers and catheters) identifications. This is likely as Imagenet pretraining allows for robust object detection and identification while CheXpert pretraining results in more condition based understanding. It is likely for this reason that combining the two approaches in the IMG+CHX:TF model results in consistently high performance across labels, increasing from CHX:TF for Support Device identification (while not quite reaching that of IMG:TF) as well as increasing from IMG:TF for Pleural Effusion (while not quire reaching that of CHX:TF). From the state of the art Chen et al paper, total F1 score of the model is reported as 0.276 which is the best in the reported literature they survey. Our F1 score of 0.433 may be increased as we use CheXbert for labeling while the Chen et al paper uses CheXpert labeling, which is significantly worse performing in labeling than CheXbert [5, 14]. Due to the differences in F1 methodology, a direct comparison of the scores across our models to the Chen et al model is discouraged.

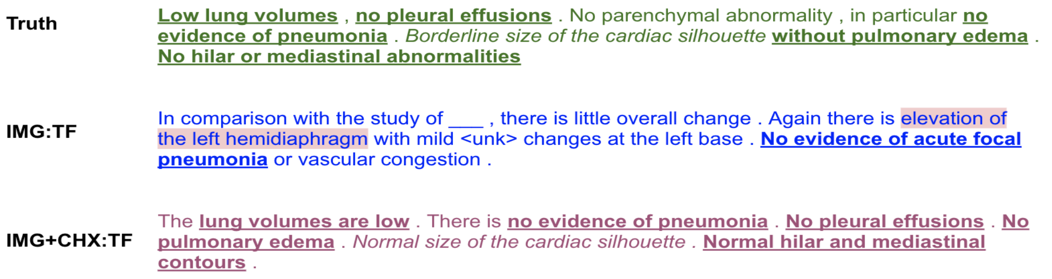
4 Analysis

In this section we seek to analyze the reports themselves that are outputs of the same input image across both the IMG:TF and IMG+CHX:TF models. We choose these two models to compare as the only difference between the two is IMG+CHX:TF being supplemented by an encoding block with CheXpert feature extraction inputs. Thereby, by doing a direct comparison between the two outputs we may illustrate the medical knowledge gained by such combined feature inputs. In Appendix Figure 5, we see an incorrect diagnosis made by IMG:TF model but No Findings reported by IMG+CHX:TF model. The IMG:TF model seems to focus less on reporting of conditions rather than more so describing what is observed. In Appendix Figure 6, IMG:TF again misdiagnosis a plural effusion which is correctly identified as absent by IMG+CHX:TF model. We see that the IMG+CHX:TF model claims the heart is mildly enlarged whereas the ground truth states that this is in fact severe. In Appendix Figure 6, an interesting effect is noted that IMG:TF identifies the pacemaker correctly and IMG+CHX:TF correctly identifies a support device; however, IMG:TF describes in more detail where the leads extend. This superior description is likely due to a better understanding of the image features due to pure Imagenet feature dependence of the model inputs. In Appendix Figure 7, again, both models correctly identify the support device and make a correct diagnosis of left pleural effusion; however, the example is included to demonstrate a problem in generating such reports. Because the ground truth reports that train the model contain a high frequency of comparison phrases to past visits, the generated reports mirror this; however, by definition the model has almost no information pertaining to past observations and thus these sentences are errant. Future approaches should seek to limit this incorrect information from making it into the final generated report. Interestingly, correct approaches to such would result in decreased BLEU(1-4) performance, again demonstrating the need to include more domain specific metrics in evaluating future models.

5 Conclusion

This project demonstrates that image captioning and report generating tasks from image or video input in the medical domain can seek to benefit from domain specific CNNs, or multimodal domain specific CNNs combined with Imagenet trained CNNs, for visual feature extraction as model inputs. Specifically, in the chest radiograph domain, Imagenet pretraining alone for the visual feature extractor results in lower performance in the pure identification of findings from the imaging. Our highest performing approach, as evaluated by total F1 score, the combined Imagenet and CheXpert pretraining should suggest future models to make use of multimodal transformers with separate encoding blocks to capture both image feature and medical specific knowledge from input image. Our work also exists as further evidence as to why BLEU score should not be used as sufficient criteria to proclaim emerging models as state of the art in the field of automated medical report generation. We demonstrate that the highest achieving BLEU score model scored the worst on F1 criteria. As we move forward in this work, we seek to use F1 labelling during training time so as to penalize or reward the loss function for medical labelling accuracy so as to increase medical knowledge of the models during their training time. Doing so may itself help the models move away from sentences describing comparisons to past presentations of the patient. Finding the right balance between cross entropy and findings label accuracy will be essential to develop automated models that have high confidence so as to be used to pave the way towards medical accessibility, medical education, and clinician assistance for error reduction.

6 Appendix



Truth

Low lung volumes , **no pleural effusions** . No parenchymal abnormality , in particular **no evidence of pneumonia** . *Borderline size of the cardiac silhouette* **without pulmonary edema** . **No hilar or mediastinal abnormalities**

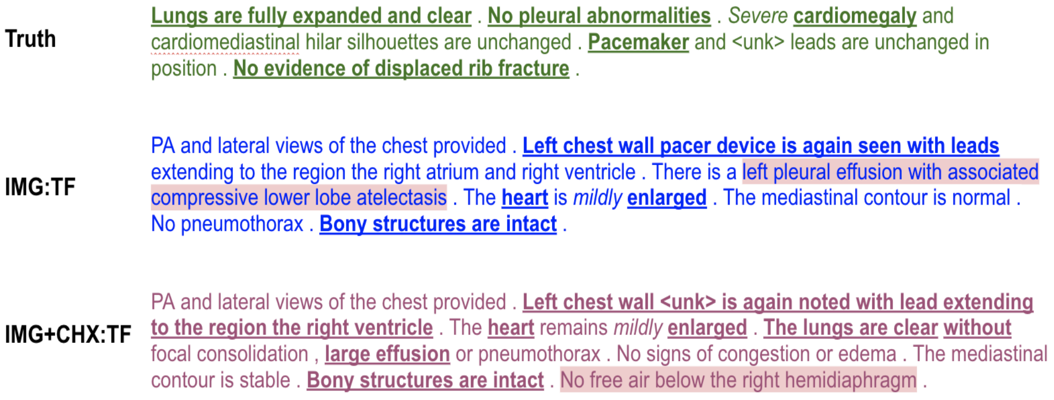
IMG:TF

In comparison with the study of ___ , there is little overall change . Again there is elevation of the left hemidiaphragm with mild <unk> changes at the left base . **No evidence of acute focal pneumonia** or vascular congestion .

IMG+CHX:TF

The **lung volumes are low** . There is **no evidence of pneumonia** . **No pleural effusions** . **No pulmonary edema** . *Normal size of the cardiac silhouette* . **Normal hilar and mediastinal contours** .

Figure 5: Generated reports in one case of "No Findings"



Truth

Lungs are fully expanded and clear . **No pleural abnormalities** . *Severe* **cardiomegaly** and cardiomediastinal hilar silhouettes are unchanged . **Pacemaker** and <unk> leads are unchanged in position . **No evidence of displaced rib fracture** .

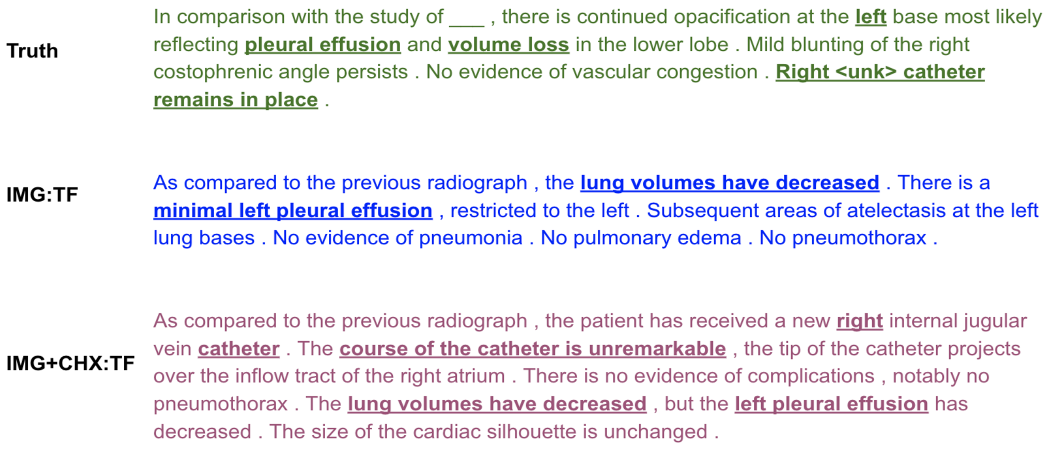
IMG:TF

PA and lateral views of the chest provided . **Left chest wall pacer device is again seen with leads** extending to the region the right atrium and right ventricle . There is a left pleural effusion with associated compressive lower lobe atelectasis . The **heart** is *mildly* **enlarged** . The mediastinal contour is normal . No pneumothorax . **Bony structures are intact** .

IMG+CHX:TF

PA and lateral views of the chest provided . **Left chest wall <unk> is again noted with lead extending to the region the right ventricle** . The **heart** remains *mildly* **enlarged** . **The lungs are clear** **without** focal consolidation , **large effusion** or pneumothorax . No signs of congestion or edema . The mediastinal contour is stable . **Bony structures are intact** . No free air below the right hemidiaphragm .

Figure 6: Generated reports in one case of "Support Devices"/"Cardiomegaly"



Truth

In comparison with the study of ___ , there is continued opacification at the **left** base most likely reflecting **pleural effusion** and **volume loss** in the lower lobe . Mild blunting of the right costophrenic angle persists . No evidence of vascular congestion . **Right <unk> catheter remains in place** .

IMG:TF

As compared to the previous radiograph , the **lung volumes have decreased** . There is a **minimal left pleural effusion** , restricted to the left . Subsequent areas of atelectasis at the left lung bases . No evidence of pneumonia . No pulmonary edema . No pneumothorax .

IMG+CHX:TF

As compared to the previous radiograph , the patient has received a new **right** internal jugular vein **catheter** . The **course of the catheter is unremarkable** , the tip of the catheter projects over the inflow tract of the right atrium . There is no evidence of complications , notably no pneumothorax . The **lung volumes have decreased** , but the **left pleural effusion** has decreased . The size of the cardiac silhouette is unchanged .

Figure 7: Generated reports in one case of "Pleural Effusion"

7 Contributions

Ethan Schonfeld was responsible for preprocessing, including: word embeddings, high frequency typo embedding corrections, and image/report parsing. Ethan Schonfeld was also responsible for feature layer extraction from the CNN pretrained models used. Ethan Schonfeld and Edward Vendrow jointly constructed the transformer models and modified their various architectures. Edward Vendrow created the dataset script and dataloader. Edward Vendrow implemented beam search. Ethan Schonfeld was responsible for BLEU implementation, CheXbert labeling, and F1 evaluation. Ethan Schonfeld wrote this report and Edward Vendrow assisted in figure construction and literature review. Dr. Greg Zaharchuk advised us that the future of chest radiograph reports will become structured text.

References

[1] William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for Chest X-Ray Report Generation. In Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 126–140. PMLR, 13 Dec 2020.

[2] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic image-based medical report generation, 2020.

[3] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, , and Pranav Rajpurkar. Chextransfer: Performance and parameter efficiency of imagenet models for chest x-ray interpretation., 2021.

[4] 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. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), 2019.

[5] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer, 2020.

[6] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. A survey on biomedical image captioning. *CoRR*, abs/1905.13302, 2019.

[7] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.

[8] Zhou HY., Yu S., Bian C., Hu Y., Ma K., and Zheng Y. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. *Martel A.L. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science*, 12261, 2020.

[9] 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. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[10] Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 457–466, Cham, 2018. Springer International Publishing.

[11] saahiluppal. Catr: Image captioning with transformers. `https://github.com/saahiluppal/catr`, 2020.

[12] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling, 2020.

[13] Jindřich Libovický, Jindřich Helcl, and David Mareček. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[14] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]