CheXbert: Labeling Chest X-Ray Radiology Reports with BERT

Stanford CS224N Custom Project

Akshay Smit
Department of Computer Science
Stanford University
akshaysm@stanford.edu

Saahil Jain
Department of Computer Science
Stanford University
saahil.jain@stanford.edu

Abstract

Using a teacher-student scheme, we train a BERT-base model to label 14 medical conditions in chest X-ray radiology reports as positive, negative, uncertain, or blank (unmentioned). Our student model, called CheXbert, is trained on a large corpus of 190,460 radiology reports from the CheXpert dataset, with labels generated automatically by the rules-based CheXpert labeler, the teacher. Despite being trained on imperfect labels, CheXbert is able to outperform or match the teacher on most medical conditions, and achieve state-of-the-art results on several tasks. CheXbert is also significantly faster than the CheXpert labeler, taking <12 seconds to label 1000 radiology reports, compared to 28 minutes for the CheXpert labeler. We further train CheXbert on a set of 1000 radiologist-labeled reports to greatly improve its generalization performance to a dataset from another institution. This suggests a general paradigm for achieveing high performance in medical settings with a paucity of human-annotated data: first train on a large set with automatically generated, imperfect labels, and then train on the smaller human-annotated data.

1 Key Information

- Mentor: Emma Chen
- External Collaborators: Pranav Rajpurkar, Anuj Pareek

2 Introduction

The chest X-ray is a rapid and critically important method to diagnose a wide range of medical conditions affecting the heart and lungs, and is globally the most widespread medical imaging examination. Automated chest X-ray interpretation can significantly improve clinical workflows, support clinical decision-making, and facilitate large-scale automated population screening. However, in order to train deep neural networks to achieve high performance on this task, hundreds of thousands of chest X-rays annotated with a wide range of pathologies are needed. Since it is infeasible for radiologists to manually annotate such a large number of X-rays, large chest X-ray datasets like CheXpert [1] and MIMIC-CXR [2] have instead used radiology reports accompanying the X-rays to automatically generate labels for over 200,000 chest X-rays. The CheXpert dataset with these automatically generated labels was released as part of a competition to encourage application of deep learning techniques to labeling chest X-ray images.

The radiology report is an unstructured textual report containing a radiologist's interpretation of the clinical findings in an X-ray. Using 1000 radiologist-annotated radiology reports, Irvin et. al. [1] developed a labeler based on hand-crafted rules to detect the presence of 14 medical conditions in CheXpert radiology reports. However, the labeler's performance was only tested on the 1000 radiology reports that were used to develop the labeler, so it's ability to generalize to other datasets and domains (such as CT scans) is suspect. Although the CheXpert labeler outperformed a previous

labeler developed by the NIH [3] on the 1000 radiologist-annotated reports, it was outperformed in many tasks by the NIH labeler on the MIMIC-CXR dataset [2]. Moreover, the rule-based labeler has no real understanding of natural language, whereas the state-of-the-art in a wide range of NLP tasks has been greatly advanced by pre-trained language representation models like BERT [4].

We fine-tune a BERT-base model, called CheXbert, to label radiology reports with 14 medical conditions. We train our model, the student, on 190,460 CheXpert radiology reports with ground truth labels provided by the rule-based CheXpert labeler, the teacher. This teacher-student training approach is inspired by computer vision, in which powerful student models have been trained on labels provided by less-sophisticated teacher models to not only outperform the teacher, but also achieve state-of-the-art results on ImageNet [5]. We evaluate this approach on the 1000-radiologist annotated reports from the CheXpert dataset, and are able to outperform the CheXpert labeler on many tasks, and match its performance on several other tasks. CheXbert also offers a 140x speedup: it takes <12 seconds to label 1000 reports (using 3 TITAN-XP GPU's) whereas the CheXpert labeler takes nearly 28 minutes. To test generalization performance to a different dataset, 150 randomly selected radiology reports from the MIMIC-CXR dataset were annotated by a radiology resident. Before evaluating on these reports, we further train our CheXbert model on the 1000 radiologist-labeled reports. This further training greatly improves generalization performance for many conditions on the 150 reports, and out of 14 medical conditions, the model is able to outperform the CheXpert labeler on 9 conditions, and match its performance on 4 conditions.

3 Related Work

Previous work in labeling chest radiology reports has centered around the use of traditional rules-based labelers. Irvin et al. [1] and Peng et al. [3] leverage dependency parsers and other traditional NLP techniques in the CheXpert labeler and the NIH NegBio tool respectively. The need for such labeling tools is apparent with the release of various large chest X-ray image datasets that rely on automated labeling of radiologist reports due to the difficulty of obtaining radiologist labels at scale. CheXpert [1], a dataset of 224,316 chest X-ray images, MIMIC-CXR [2] a dataset of 377,110 chest X-ray images, and ChestX-ray8 [6], a dataset of 108,948 X-ray images, all obtain image labels from radiology reports using rules-based natural language processing. PadChest [7], a Spanish dataset of more than 160,000 images, uses recurrent neural networks with attention mechanisms to label 73 percent of their images.

None of the automated labelers across these various projects take advantage of transformers and advanced language representation models like BERT [4]. In addition to BERT, we also examine the usefulness of BioBERT [8], which further pre-trains BERT weights on medical data, on our label extraction task. To train a label extractor, we treat the CheXpert labeler as a teacher and our BERT/BioBERT-based model as a student. Xie et al. [5] show that such a student-teacher paradigm can achieve state of the art results in vision.

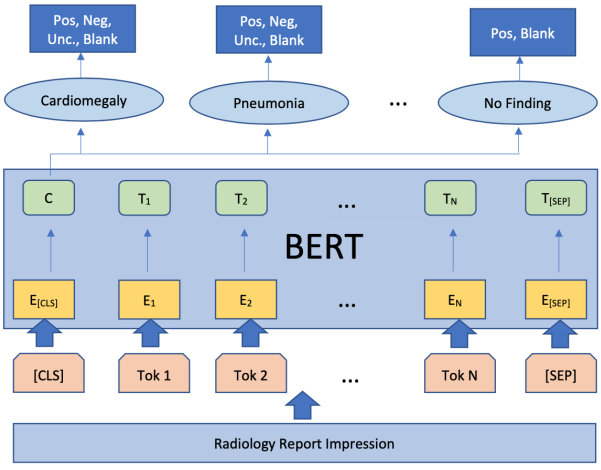

Figure 1: The CheXbert labeler architecture

Condition	Positive	Negative	Uncertain	Blank
Atelectasis	29,818 (15.66%)	1,018 (0.53%)	29,832 (15.66%)	129,792 (68.15%)
Cardiomegaly	23,302 (12.23%)	7,809 (4.10%)	6,682 (3.51%)	152,667 (80.16%)
Consolidation	12,977 (6.81%)	19,397 (10.18%)	24,345 (12.78%)	133,741 (70.22%)
Edema	49,725 (26.11%)	15,867 (8.33%)	11,746 (6.17%)	113,122 (59.39%)
Enlarged Cardiomediastinum	9,129 (4.79%)	15,165 (7.96%)	10,278 (5.40%)	155,888 (81.85%)
Fracture	7,364 (3.87%)	1,960 (1.03%)	488 (0.26%)	180,648 (94.85%)
Lung Lesion	6,955 (3.65%)	758 (0.40%)	1,084 (0.57%)	181,663 (95.38%)
Lung Opacity	94,156 (49.44%)	5,006 (2.63%)	4,404 (2.31%)	86,894 (45.62%)
No Finding	16,795 (8.82%)	NA	NA	173,665 (91.18%)
Pleural Effusion	77,028 (40.44%)	25,097 (13.18%)	9,565 (5.02%)	78,770 (41.36%)
Pleural Other	2,481 (1.30%)	210 (0.11%)	1,801 (0.95%)	185,968 (97.64%)
Pneumonia	4,647 (2.44%)	1,851 (0.97%)	15,907 (8.35%)	168,055 (88.24%)
Pneumothorax	17,688 (9.29%)	47,566 (24.97%)	2,704 (1.42%)	122,502 (64.32%)
Support Devices	107,601 (56.50%)	5,319 (2.79%)	910 (0.48%)	76,630 (40.23%)

Table 1: Class prevalence for each medical condition on the CheXpert dataset, comprising both our train and dev sets, giving a total of 190,460 reports.

4 Approach

To our knowledge, we are the first to apply sophisticated pre-trained language representation models like BERT to chest radiology reports. The baseline we compare our model against is the rules-based CheXpert labeler developed by Irvin et. al. [1]. We wrote all the code for this project ourselves, except for a script provided by Hugging Face to convert Tensorflow checkpoints to PyTorch, and the code to pre-train the BERT weights on our radiology corpus. All our deep models were implemented in PyTorch.

4.1 CheXbert model

CheXbert consists of a pre-trained BERT-base model with 14 linear heads, one for each medical condition (see Figure 1). We extract and tokenize the "Impression" section of each radiology report, and feed it as input into the BERT model. The maximum number of tokens in each input sequence is capped at 512, and sequences longer than this are cut off after the 512th token. Only 3 reports in the CheXpert dataset are longer than 512 tokens. The Impression section is a summary of key clinical findings in the radiographic study. The final-layer's hidden state corresponding to the CLS token is then fed as input to each of the linear heads. 13 of the linear heads output 4 class scores, corresponding to the positive, negative, uncertain and blank classes. Of these 13 heads, 12 correspond to various pathologies, and 1 corresponds to medical support devices. The 14th head corresponds to "No Finding", for which there are only 2 classes: positive and blank.

4.2 BERT vs. BioBERT

While the original BERT model was pre-trained on a large general text corpus including English Wikipedia, Lee et. al. [8] further tuned the BERT weights by pretraining on a large medical corpus. We therefore apply the linear heads on top of BERT vs. on top of BioBERT, and compare the performance of the two approaches after fine-tuning on our medical condition detection task. Surprisingly, we find plain BERT to significantly outperform BioBERT on our tasks.

4.3 Pretraining BERT on radiology report corpus

We further experiment with pre-training the BERT weights on our radiology report corpus, and then fine-tuning on the medical condition detection task. The pretraining corpus consists of the "Impression" sections of 377,134 radiology reports from both the CheXpert and MIMIC-CXR datasets, for a total of 12,122,090 words.

4.4 Freezing BERT's embedding layer

In another experiment, we freeze the embedding layer of the BERT model and then perform fine-tuning. Since the total corpus of our radiology reports is considerably smaller than the corpus used to

pretrain BERT, our corpus may not cover most of the words in BERT's vocabulary. So fine tuning the embedding layer might shift the weights for some portion of the vocabulary, while leaving out synonyms or other related words, leading to a worse embedding.

4.5 Up-sampling for rare classes

For many of our medical conditions, 'uncertain' and 'negative' labels are rare (see Table 1). For instance, only 1.42% of reports had an 'uncertain' label for Pneumothorax. The rarity of these labels causes our model to perform noticeably worse on detecting negation and uncertain labels for these conditions. We therefore up-sample reports with rare class labels in the following manner.

Let $\mathbb{R}$ be the set of all CheXpert radiology reports with labels provided by the rule-based CheXpert labeler. Let $\mathbb{C}$ be the set of 13 medical conditions, excluding No Finding. And let $\mathbb{L} = \{positive, negative, blank, uncertain\}$ be the set of all class labels. Then for a report $r \in \mathbb{R}$ and condition $c \in \mathbb{C}$, let $r[c] \in \mathbb{L}$ be the class label for condition c in report r. For a label $l \in \mathbb{L}$, we define the weight

$$w_c[l] = \frac{|\mathbb{R}|}{\sum_{r \in \mathbb{R}} \mathbf{1}\{r[c] == l\}}$$

where $\mathbf{1}\{r[c] == l\}$ is the indicator function which equals 1 when $r[c] == l$, else 0. $w_c[l]$ should be interpreted as the weight given to label l for the condition c, with rarer labels getting a higher weight. We then define the weight given to report r as:

$$w[r] = \sum_{c \in \mathbb{C}} w_c[r[c]]$$

Thus, reports which contain rarer labels for more medical conditions get higher weight. Note that we had per-medical condition weights $w_c[l]$ for each class label, since labels that are very rare for one medical condition may not be nearly as rare for another condition. The weights $w[r]$ are normalized to produce probabilities $p[r]$. During fine-tuning, we create batches though random sampling with replacement, with report r being picked with probability $p[r]$. For No Finding, the weight calculations are the same as above, except that No Finding can only have the labels 'positive' or 'blank'.

4.6 Focal loss for poorly-classified examples

Lin et al. [9] introduced the focal loss function, which modifies the regular cross-entropy loss to emphasize poorly-classified examples, and de-emphasize examples on which the model does very well. We use the focal loss as an alternative to the cross-entropy loss in order to improve performance on rare class labels, on which the model is likely to do poorly. For a fixed report r and fixed condition c, let $p_t(r, c)$ be the post-softmax probability given by the model to the ground truth label for condition c. Then we define

$$FL(r, c) = -(1 - p_t(r, c))^\gamma \log(p_t(r, c))$$

where γ is a hyperparameter. The focal loss for the report r is then defined as

$$FL(r) = \sum_{c \in \mathbb{C}} FL(r, c)$$

Note that $FL(r, c)$ simply multiplies the regular cross-entropy loss by a factor of $(1 - p_t(r, c))^\gamma$. For well classified examples, $p_t \to 1$ so that this factor is small and the loss for that example is down-weighted. Conversely the loss for a poorly-classified example is up-weighted.

4.7 Further training on 1000 radiologist labels

After training several models on the large CheXpert dataset, we evaluated their performance on a separate set of 1000 radiologist-annotated reports. Subsequently, we further train our best-performing model on these 1000 reports, and refer to this model as CheXbert + 1000. We evaluate the performance of CheXbert + 1000 on 150 reports from the MIMIC-CXR dataset (manually labeled by a radiology resident). Our first goal is to evaluate our model's generalization performance to data from another institution. Our second goal is to determine if a model trained on a large dataset with imperfect labels can benefit from further training on a small dataset with high-quality labels. Surprisingly, further training on the 1000 radiologist-labeled reports significantly increases performance for certain conditions on the 150 MIMIC-CXR reports.

	BioBERT	BERT	BERT+freeze	BERT+pretrain	BERT+sampling	BERT+FL
Mention-F1 avg.	0.946	**0.948**	0.947	**0.948**	0.947	0.947
Negation-F1 avg.	0.793	0.907	0.909	**0.916**	0.897	0.828
Uncertain-F1 avg.	0.597	**0.748**	0.735	**0.748**	0.747	0.733

Table 2: F1 scores averaged over all 14 conditions. FL is focal loss. BERT+pretrain is pretrained on the combined CheXpert and MIMIC-CXR corpus.

5 Data

5.1 CheXpert Dataset

To train our model, we use the CheXpert dataset compiled by Irvin et al. [1], which contains 224,316 chest radiographs of 65,240 patients collected between October 2002 and July 2017 from Stanford Hospital. The reports associated with X-rays were labeled for the presence of the following 14 common chest conditions: Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices, and No Finding. While No Finding has two possible labels of positive or blank, each of the other 13 conditions have four possible labels: positive, negative, uncertain, or blank. Positive indicates the presence of a condition, negative indicates the absence of a condition, uncertain indicates that the condition may or may not be present, and blank indicates that the condition was not mentioned. Table 1 shows the class prevalence for each medical condition.

The input to our CheXbert model is a tokenized report impression, which is a summary of clinical findings in a radiology report. An example report impression is: "A frontal and lateral study of the chest shows a midline trachea. The heart, aorta and remaining visualized mediastinal structures appear unremarkable." After removing duplicate reports in the CheXpert dataset (as duplicate records exist for frontal and lateral views of the chest) and removing the 1000 radiologist labeled reports (which function as our test set), we obtain 190,460 report impressions. Using an 85-15 percent train-dev split, our train, dev, and test sets ultimately consist of 161891, 28569, and 1000 report impressions respectively. Note that only our test set has radiologist labels corresponding to its reports; the train and dev sets both only have labels automatically generated by the CheXpert labeler.

5.2 MIMIC-CXR Dataset

We obtained access to reports from the MIMIC-CXR dataset [2] created by MIT. Although the creators of this dataset used the same medical conditions and labels as CheXpert, we only obtained the raw, unlabeled radiology reports in MIMIC-CXR. We clean and extract report impressions from 150 MIMIC-CXR reports, which are labeled manually by a radiology resident to construct another test set to check generalization of our model to a dataset from a different institution. We also extract a set of 186,674 report impressions from MIMIC-CXR (not overlapping with the 150 report test set) to further pretrain BERT weights on radiology report impressions.

6 Experiments

6.1 Evaluation method

We use a number of distinct evaluation tasks to measure the performance of our model and the baseline on detecting negations, uncertains, and mentions of medical conditions. We also evaluate inter-observer agreement between the ground truth labels and our model's predictions.

6.1.1 Mention-F1, Negation-F1 and Uncertain-F1

In the mention detection task, we treat every label except 'blank' as positive, and 'blank' labels are treated as negative. The aim of this task is to evaluate the model's ability to determine if a condition was mentioned in the report or not. With this binarization of labels, we compute a F1-score for each condition, called the Mention-F1 score. The negation and uncertain detection tasks are similar. In the negation task, only 'negative' labels are treated as positive, and in the uncertain task, only 'uncertain' labels are treated as positive. The resulting F1 scores for each condition are called the Negation-F1 and Uncertain-F1 scores. Note that these three tasks are the same as in the CheXpert paper [1].

	Mention F1				Negation F1				Uncertain F1			
Category	n	NIH	CheXpert	Ours	n	NIH	CheXpert	Ours	n	NIH	CheXpert	Ours
Atelectasis	301	0.976	**0.998**	**0.998**	5	0.526	**0.833**	**0.833**	152	0.661	0.936	**0.950**
Cardiomegaly	203	0.647	0.973	**0.975**	69	0.000	0.909	**0.933**	17	0.211	**0.727**	0.407
Consolidation	367	0.996	**0.999**	**0.999**	187	0.879	**0.981**	**0.981**	123	0.438	0.924	**0.929**
Edema	354	0.978	**0.993**	**0.993**	106	0.873	0.962	**0.967**	55	0.535	0.796	**0.818**
Pleural Effusion	538	0.985	**0.996**	**0.996**	195	0.951	0.971	**0.974**	41	0.553	0.707	**0.714**
Pneumonia	126	0.660	**0.992**	**0.992**	16	0.703	0.750	**0.774**	81	0.250	0.817	**0.838**
Pneumothorax	384	0.993	**1.000**	**1.000**	327	0.971	**0.977**	0.975	10	0.167	**0.762**	0.737
Enlarged Cardiom.	212	N/A	**0.935**	**0.935**	125	N/A	0.959	**0.963**	51	N/A	**0.854**	0.839
Lung Lesion	48	N/A	**0.896**	**0.896**	11	N/A	**0.900**	**0.900**	8	N/A	**0.857**	**0.857**
Lung Opacity	472	N/A	**0.966**	**0.966**	34	N/A	**0.914**	**0.914**	5	N/A	0.286	**0.316**
Pleural Other	42	N/A	**0.850**	0.835	1	N/A	**1.000**	**1.000**	12	N/A	**0.769**	**0.769**
Fracture	80	N/A	**0.975**	**0.975**	32	N/A	0.807	**0.900**	6	N/A	**0.800**	**0.800**
Support Devices	447	N/A	**0.933**	0.930	23	N/A	0.720	**0.792**	0	N/A	N/A	N/A
No Finding	119	N/A	0.769	**0.780**	0	N/A	N/A	N/A	0	N/A	N/A	N/A
Average		N/A	**0.948**	**0.948**		N/A	0.899	**0.916**		N/A	**0.770**	0.748

Table 3: F1 scores across all conditions and tasks on 1000 radiologist-labeled reports. n is the number of positive examples.

	Cohen's Kappa	
Category	CheXpert	Ours
Atelectasis	0.953	**0.964**
Cardiomegaly	**0.928**	0.865
Consolidation	0.962	**0.963**
Edema	0.949	**0.955**
Pleural Effusion	0.947	**0.948**
Pneumonia	0.852	**0.865**
Pneumothorax	**0.971**	0.969
Enlarged Cardiom.	**0.884**	0.882
Lung Lesion	**0.860**	**0.860**
Lung Opacity	0.913	**0.916**
Pleural Other	**0.794**	0.778
Fracture	0.907	**0.940**
Support Devices	**0.865**	**0.865**
No Finding	0.737	**0.745**

Table 4: κ scores across all conditions and tasks on 1000 radiologist-labeled reports.

	Cohen's Kappa		
Category	CheXpert	CheXbert	CheXbert + 1000
Atelectasis	0.856	**0.893**	**0.910**
Cardiomegaly	0.754	0.698	**0.773**
Consolidation	**0.950**	**0.950**	**0.950**
Edema	**0.898**	0.884	0.884
Pleural Effusion	0.889	**0.903**	**0.903**
Pneumonia	0.715	0.715	**0.750**
Pneumothorax	0.942	**0.981**	**0.981**
Enlarged Cardiom.	0.560	**0.674**	**0.674**
Lung Lesion	**0.322**	**0.322**	**0.322**
Lung Opacity	0.764	**0.765**	0.764
Pleural Other	**0.664**	**0.664**	**0.664**
Fracture	0.709	0.709	**0.855**
Support Devices	0.870	0.870	**0.902**
No Finding	0.708	0.726	**0.754**

Table 5: κ scores on 150 MIMIC-CXR reports.

6.1.2 Cohen's Kappa

We use the Cohen's Kappa $\kappa \in [-1, 1]$ to measure inter-observer agreement between the ground truth labels and the model's predictions. $\kappa = 1$ indicates perfect agreement, whereas values below 0 indicate random or worse agreement. Unlike accuracy, κ accounts for chance agreement that would occur if the two observers were to assign labels randomly. We define

$$\kappa = \frac{p_{acc} - p_c}{1 - p_c}$$

where p_{acc} is simply the accuracy, and p_c is the probability of chance agreement. For further details, see [10].

6.2 Experimental details

In all our experiments, models were trained for 4 epochs using Adam optimization with a learning rate of 2×10^{-5}, in line with the recommendations of Devlin et. al. for fine-tuning tasks [4]. We found our models to converge within 4 epochs. Attempts to use SGD with momentum, or annealing the learning rate, did not improve performance. During training, we periodically compute Cohen's

Kappa κ for each condition on our dev set, average over the conditions, and save the checkpoint with the highest average. All models were trained using 3 TITAN-XP GPU's with a batch size of 18, as larger batches would not fit in GPU memory. In the focal loss experiment, the highest performance was achieved with $\gamma = 2$.

6.3 Results

6.3.1 BioBERT vs. BERT-based models

Table 2 shows the averaged F1 scores of several of our models on the 1000 radiologist-annotated CheXpert reports. For each of the models, we compute the Mention-F1, Uncertain-F1 and Negation-F1 scores for all 14 conditions as described above. We then average these F1 scores over all conditions to produce the average F1 scores. The BERT model pre-trained on CheXpert + MIMIC-CXR and fine-tuned with vanilla cross-entropy loss outperforms all other models, although its performance is close to that of the non-pretrained BERT model. Up-sampling rare classes, using focal loss, or freezing the embedding layer does not improve performance. Most surprisingly, the BERT weights significantly outperform the BioBERT weights even though BioBERT was additionally pre-trained on a large biomedical corpus. For all subsequent experiments, the terms 'CheXbert' and 'Ours' refer to the BERT+pretrain model.

6.3.2 NIH vs. CheXpert vs. CheXbert

Table 3 shows the F1 scores for the NIH labeler, CheXpert labeler, and CheXbert across all conditions and tasks, on the 1000 radiologist-labeled reports. The average is simply the average of the F1 scores across all conditions. CheXbert appears to very closely mimic the performance of the CheXpert labeler on the mention task, with both models obtaining identical scores for most conditions. CheXbert offers the greatest improvement on negation detection: it matches or outperforms the CheXpert labeler on all but one negation task, with a higher negation-F1 average. Although CheXbert's uncertain-F1 average is lower than that of CheXpert, the CheXpert labeler only outperforms CheXbert on 3 conditions (Cardiomegaly, Pneumothorax and Enlarged Cardiomediastinum). CheXbert's uncertain-F1 average is particular lowered by its poor performance on Cardiomegaly, which contained only 17 uncertain labels across all 1000 reports. More uncertain labels for such conditions would be required to conclude a significant difference in performance between CheXpert and CheXbert on the uncertain task.

Poor performance on a few conditions having a dearth of uncertain labels can greatly reduce the uncertain-F1 average. Therefore, we also measure inter-observer agreement between radiologists and CheXbert/CheXpert using the κ metric in Table 4. CheXbert outperforms CheXpert on 8 conditions, and matches its performance on 2 conditions. Both models achieve very similar κ scores on several tasks, which again suggests that CheXbert has learned to closely mimic the teacher, while still managing to surpass the teacher on several conditions.

6.3.3 Generalizing across institutions with CheXbert + 1000

Since the CheXpert labeler was developed and tested on the same set of 1000 radiologist-labeled reports, its generalization performance is suspect. Indeed, it's performance on the MIMIC-CXR dataset was considerably worse than on the 1000 reports used to develop the labeler [2]. We randomly selected 150 reports from MIMIC-CXR, and these were manually labeled by a radiology resident. Before evaluating on these 150 reports, we further train CheXbert on the 1000 radiologist-labeled reports, calling this model CheXbert + 1000. The κ scores on the 150 MIMIC-CXR reports are given in Table 5.

Surprisingly, further training on a small set with high-quality labels allows CheXbert+1000 to outperform CheXpert on 9 tasks and match its performance on 4 tasks. Further training on the 1000 reports also offers significant performance gains for several conditions, such as Cardiomegaly, Pneumonia, Fracture, Support Devices and No Finding. Furthermore, CheXbert only outperforms CheXbert + 1000 on Lung Opacity, and only by a tiny margin. Training on the 1000 reports is most useful after the model has already been trained on a large corpus with possibly imperfect labels. We trained a BERT model directly on the 1000 radiology reports, without training on our large train set of CheXpert reports. This model performed poorly, getting κ scores near 0 for many conditions.

6.3.4 Cross-domain transfer to CT scans

To test and analyze CheXbert on another task in a different domain other than chest X-rays, we obtained two manually-labeled report impressions from chest CT scans. The first example report impression was the following:

"1. Patchy right upper lobe and lingula lobe consolidation and groundglass which are suspicious for pneumonia. 2. Main pulmonary artery enlargement which may be related to pulmonary hypertension. There are no substantial differences between the preliminary results and the impressions in this final report."

CheXbert correctly predicted consolidation as positive and pneumonia as uncertain, most likely taking advantage of a semantic understanding of the word "suspicious." The second example report impression was the following:

"1. Status post replacement of the ascending aorta with continued healing of the sternotomy and decrease in minimal residual postsurgical changes in the anterior mediastinum. 2. Few stable nodules in the lung measuring up to 4 mm and improved linear atelectasis."

CheXbert correctly labeled atelectasis and lung lesion as positive but mistakenly labeled enlarged cardiomediastinum as positive instead of blank. Despite never having seen a report from CT scans, which are very different from X-rays, CheXbert is still able to generalize well on these two reports.

7 Conclusion

We train a BERT-base model with linear heads as a student model trained on a large corpus of radiology reports, with labels provided by the rule-based CheXpert labeler, the teacher. For some of our 14 medical conditions, uncertain and negative labels are quite rare, and state-of-the-art neural networks are known to require large amounts of data to achieve high performance. Despite this, the student model (CheXbert) is able to closely mimic the teacher, matching or surpassing the teacher's performance on most tasks on a set of 1000 radiologist-labeled reports. BERT weights significantly outperformed BioBERT weights, which is surprising since the latter was further trained on a large biomedical corpus. We achieve highest performance by further pre-training the BERT weights on our large radiology corpus containing both CheXpert and MIMIC-CXR reports.

We further train CheXbert on 1000 radiologist-labeled reports (CheXbert + 1000) and evaluate generalization performance to data from another institution. For this, 150 randomly selected reports from MIMIC-CXR were manually annotated by a radiology resident. CheXbert + 1000 achieves considerable performance gains over CheXbert in many of our 14 medical conditions, outperforms CheXpert on 9 conditions and matches CheXpert on 4 conditions. Overall, CheXbert shows promising generalization performance on cross-institutional data.

CheXbert+1000 shows that models trained on a large corpus with imperfect labels can greatly benefit by being further trained on a small set with high-quality labels. This is particular pertinent in medical settings where medical specialist-level annotations are scarce. If hand-crafted rules and heuristics can be used to produce a large set of imperfect labels, then large, sophisticated models like BERT could achieve state-of-the-art performance by first training on the automatically generated labels, and then training on a smaller set of human-annotated examples. The order appears to be important, as training a BERT model on only the 1000 radiologist-labeled reports led to poor performance.

In addition to achieving state-of-the-art results for many tasks across several medical conditions, CheXbert offers 140x lower inference-time latency, taking less than 12 seconds to label 1000 reports (using 3 TITAN-XP GPU's), compared to 28 minutes for the CheXpert labeler. At scales of hundreds of thousands or millions of reports, the latency of CheXpert will be prohibitive.

For future work, we will obtain more manually-annotated MIMIC-CXR reports so as to better assess the generalization performance of our model. We shall also examine cross-domain performance on a set of manually-labeled CT scan reports, without any training on CT scan reports. Furthermore, since models like multilingual BERT have shown promising results in zero-shot transfer across languages for many tasks [11], we shall train a multilingual model and perform zero-shot transfer to reports in a different language.

References

[1] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[2] Alistair E W Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[3] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

[6] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[7] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports, 2019.

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[10] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, Oct 2012.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert?, 2019.