Predicting Respiratory Illness using Chest X-rays

Sam Spinner, Ryan Cao, Kiara Nirghin [*]
Department of Computer Science
Stanford University

Abstract

Chest X-rays are cheap, accessible, and commonly used as first tests in diagnosing many lung diseases. Many hospital systems have vast amounts of X-ray data that can be leveraged by AI systems to supplement radiologist predictions and improve patient outcomes. In particular, radiologists pore over large numbers of x-rays, CT scans, and MRI images every day [7], implying that a detection framework which can serve as a pre-filter system, drawing bounding boxes around likely areas of disease, would both significantly reduce a radiologist's workload and help them ignore negatives and hone in on positives, increasing their own diagnostic capacity. To this end, we train both a multi-label classifier (ResNet-18) and a diseased-area detector (Faster R-CNN + MobileNet backbone) on the NIH Kaggle dataset, and are able to achieve 95% validation Acc_{IoU} (defined later) with our classifier and pixel IoU of .206 with our object detection model.

1 Introduction

Chest X-rays are a staple tool for preliminary diagnosis of many respiratory diseases. As such, building an ML model to predict diseases based on X-ray could be extremely beneficial. This was evident in the Covid-19 pandemic, where increased vigor within the medical and AI fields resulted in numerous chest X-ray Covid-19 classification models being developed. While initial results seemed promising [8], an analysis of numerous approaches showed that seemingly impressive Covid-19 detection models were learning data artifacts instead of actual medical pathology [1]. Out of a set of 2,212 Covid-19+ML studies, 415 of which were examined post screening and 62 closely reviewed, *zero models were identified as clinically useful due to underlying flaws* [10]. These recent undesirable outcomes have highlighted the need for explainable, interpretable machine learning as a prerequisite for clinical deployment across many disparate fields of medicine.

Healthcare is a data-rich field, but the data is often complex and requires expert-level annotation to interpret. For this reason, many potential problems in healthcare lack sufficient data for a machine learning model to train on [2]. For problems where a sufficient dataset exists, the interaction between care providers and a trained model is still uncertain. Promising examples from diabetic retinopathy screening show that optometrists and machine learning models perform better working in concert than either alone [6], but this is a limited application example without lethal failure modes. In situations where a mis-classification by a machine learning system is the difference between life and death, extensive work is still needed to display why a model reached the decision that it did. Attention-based approaches to classifying diseases via X-ray imagery like in [5] are promising, and can display bounding boxes or heat maps to display to experts what the ML model sees. We intend for our project to make an impact in the interpretatability space.

For the classification component of this project, we have that

[*] Special thanks to our project TA, Ayush Kanodia, for providing us with bountiful insights and assistance.

CS230: Deep Learning, Fall 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

- The input to our algorithm is a re-scaled 224 x 224 chest x-ray image.
- The output labels (for classification) are one-hot class labels, with potentially multiple labels to one image (e.g. an x-ray might be classified as both having emphysema and containing a small nodule on its right side).
- The output boxes (for object detection) are given as (x, y, w, h) tuples for about 1000 of the images, and represent bounding boxes around which the disease in question can be found.
- As stated earlier (and will be described in the methods section), we used a ResNet-18 for classification and a Faster R-CNN + MobileNet backbone for object detection. The outputs of the ResNet are per-class probabilities, and the outputs of the Faster R-CNN are bounding box proposals with confidence scores attached.

2 Related Work

ChestX-Ray8

This work [11], which introduced the NIH dataset of 100k chest x-ray images, provided a strong classification baseline/inspiration. Their key approach involved taking a DenseNet, removing the final few FC and classification layers, and replacing them with a transform layer (effectively a size normalization layer which converts the feature maps into a standard size foor the classification layer). Moreover, they experimented with various loss functions and eventually settled on a weighted binary cross-entropy loss:

$$-\beta_P \sum_{y_c=1} \log(f(x_c)) - \beta_N \sum_{y_c=0} \log(1 - f(x_c)) \tag{1}$$

Where β_P, β_N represent the relative proportions of positive/negative examples within the dataset.

LSTM Label-Dependent Diagnosis

This subsequent work [12] sought to take advantage of and explicitly model the dependencies between various labels (e.g. the presence of atelectasis makes it more likely that a particular lung is also diagnosed with pneumonia), and did so by using an LSTM to output the final prediction labels one-by-one – in other words, the image features from the main convnet were fed into an LSTM for exactly c "timesteps" (where c is the number of potential classes present within an image), with each (ith) "timestep" outputting the probability that the ith class is present within the image.

CheXNet

This work tackled the classification task using a much deeper (121 layer) DenseNet, showing that a larger network genuinely increases task performance – note that the training procedure had almost no extra bells and whistles, yet achieved near-human level performance (and in fact, exceeded radiologists' performances on pneumonia); the main contribution here (aside from their state-of-the-art F1 scores) is their activation map visualizations, which effectively show which parts of the input image contribute most to the classification score computation made at the final layer, indeed verifying that the model is finding the correct visual features/regions correlated with the presence of diseases.

3 Dataset and Features

The dataset we have chosen was provided by the National Institutes of Health and is comprised of 112,120 Chest X-ray images from 30,805 unique patients. Each X-ray image is of size 1024x1024 pixels and has a single channel. The creators of the dataset used Natural Language Processing to text-mine disease labels from the radiological reports associated with each image. There are 15 label classes– 14 diseases and asymptomatic lungs. Images can be classified as "No findings" or one or more disease classes including: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule Mass and Hernia. Because they were generated by an NLP algorithm, the labels are expected to be $>90\%$ accurate and suitable for weakly-supervised learning. Additionally, — diseased images have

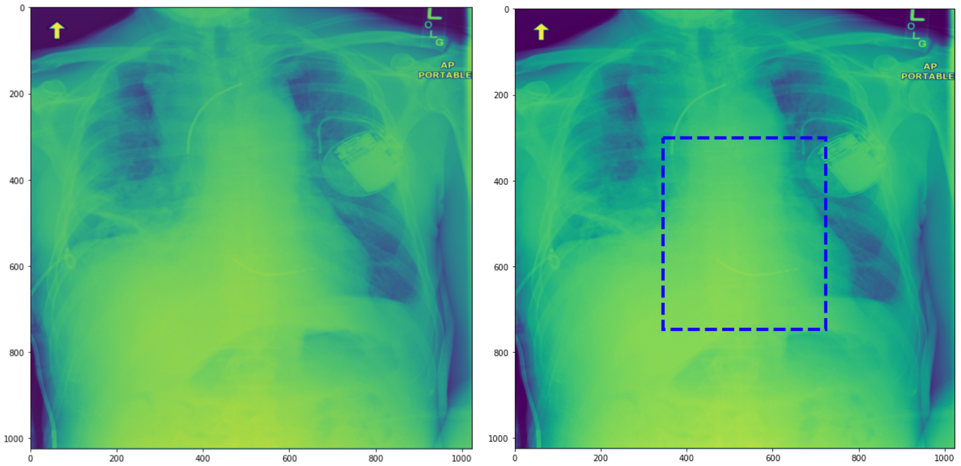


Figure 1: X-ray images labeled 'cardiomegaly'. Each image is of shape 1024x1024 with a single channel. A bounding box annotation is provided for the right image.

bounding-box annotations that were drawn by a radiologist. These bounding boxes are represented in the dataset by the bottom-right coordinate, the height, and the width.

For the classification task, we created a train set of — images and a test set of — images.

For the object detection task, our train dataset contained 801 images and our validation dataset contained 113 images.

We pre-processed training and validation images in both tasks by normalizing with the dataset-wide pixel mean and standard deviation. For the classification task, images were down-sampled to size

4 Methods

Multi-Label Classification

Model + Training – We used a fresh (non-pretrained) ResNet-18 model, modified to output a 15-dimensional vector of probabilities (no softmax, but rather per-class sigmoid activations). Note that ResNet was first described by Kaiming He et. al. at FAIR back in 2015, and involves providing "residual" identity connections between convolutional blocks.

Our loss function was a weighted, per-class binary cross-entropy loss (positive examples were weighted inversely proportionally to their frequency within the dataset) – see equation (1) above.

We used Adam as our optimizer at a learning rate of 3e-4.

Evaluation – We created a custom "IoU" accuracy metric for multi-label classification. To illustrate this metric, let $y, \hat{y} \in \{0,1\}^k$ be the ground truth and predicted labels for each of the disease classes, respectively. Then we have that

$$\text{Acc}_{\text{IoU}} = \frac{y^\top \hat{y}}{\sum_{i=1}^{k} \max(y_i, \hat{y}_i)} \tag{2}$$

Intuitively, we have that Acc_{IoU} is simply the number of correct labels over the total number of classes which were correct or predicted, or both.

Moreover, we defined a custom confusion matrix, defined thus – again, let $y, \hat{y} \in \{0,1\}^k$ be the ground truth and predicted labels for each of the disease classes, and let $M \in \mathbb{N}^{k \times k}$ be our confusion matrix. Moreover, let $\hat{y}_{inc} \in \{0,1\}^k$ be the "incorrect" labels, i.e. $\hat{y}_{inc}$ is 1 wherever $\hat{y}$ is 1 and y is 0. Then for each index i where $y_i = 1$ we have that

$$M_i := M_i + \frac{1}{\sum_{j=1}^{k} y_j} \hat{y}_{inc} + \hat{y}_i \tag{3}$$

In other words, each "incorrect" label contributes $\frac{1}{c}$ (where c is the number of correct classes) to each row of the confusion matrix corresponding to a "correct" class. If there is a single correct label, the above formula reduces to the canonical confusion matrix computation.

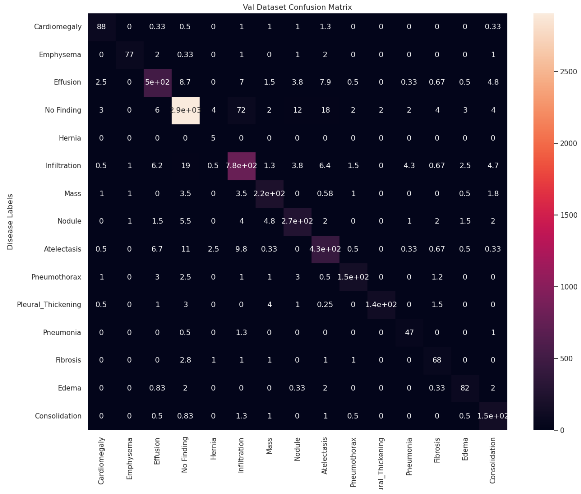
Bounding Box Object Detection

We used a fresh (non-pretrained) MobileNet architecture as the backbone of our Faster R-CNN pipeline for object detection. Note that MobileNet is an efficient convolutional architecture which relies on separating out "pointwise" $1 \times 1 \times c$ and "depthwise" $w \times h \times 1$ convolutional kernels, rather than have a "full $w \times h \times c$ kernel.

Faster R-CNN [9] is an end-to-end pipeline for object detection. In summary, the technique is the latest evolution in the R-CNN [4] series; the intuition behind the primary network in the RCNN series is that we begin with an input image and eventually produce image patch features (in Fast R-CNN [3], the patch features are computed by caching a feature computation over the entire image and then segmented appropriately, rather than computing a forward pass over each raw image patch), which are then classified into $k + 1$ classes, the extra class being a discarded "background" class. Finally, all three architectures feature a bounding box "correction" step, which learns (via regression) parameters adjusting the center and side lengths of the patches, producing the final bounding boxes.

5 Experiments/Results/Discussion

5.1 Classification Performance



The confusion matrix for the classifier (trained after 60 epochs) on the validation set. Indeed, the classifier gets an Acc_{IoU} (defined above) of 95% on the val set after training for just 30 epochs! Loss/Acc_{IoU} curves are omitted for brevity.

5.2 Object detection performance

Our Faster-RCNN model trained for 50 epochs on the bounding-box labeled dataset achieved an average pixel-IoU of .206. This means that on average, a quarter of the area of the model's predicted bounding boxes overlapped with the true bounding box. The average-pixel IoU increased with training, showing that the model was learning to identify some diseases. However, the labeled data was too small and covered too many diseases for the model to perform at a level that could be used in a clinical setting. This training disappointment emphasizes the need for radiologists to annotate more X-ray data and the important of developing semi or self-supervised object detection methods.

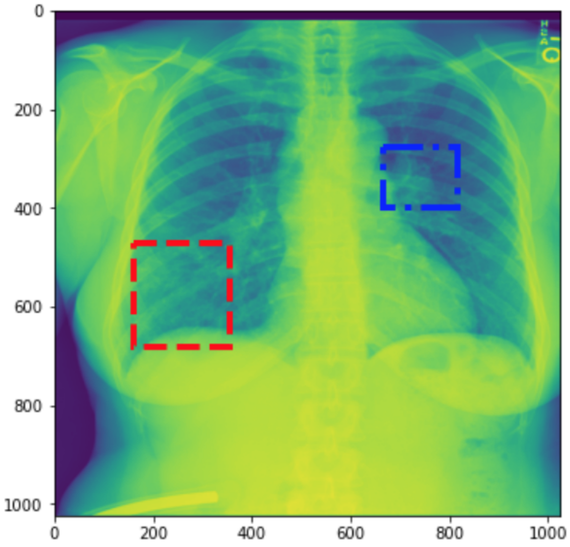


Figure 2: The red, left-most bounding box was predicted by the trained Faster-RCNN teacher model, whereas the right-most blue box is the true bounding box of the disease. It appears that the model sometimes classifies breasts and nipples as diseased regions, suggesting the need for a biological sex field to be included with each X-ray image.

5.3 The effect of post-nms filtering values

Due to our Faster-RCNN being trained on a relatively small amount of data, it seemed to predict many more detections than were actually present in the validation set, which had one detection per X-ray. We tried to mitigate this behavior in two ways. First, because the model returns a classification score with each prediction, we filtered generated predictions to only use the bounding box associated with the highest classification score. Additionally, we attempted to limit the number of region proposals that were passed to the final network in the Faster-RCNN pipeline.

5.4 The potential need for a biological sex label

We noticed that our teacher Faster RCNN model was oftentimes drawing bounding boxes around what appeared to be biological women's nipples (Fig 2), even if the bounding box label was elsewhere in the chest X-ray. Upon further examination, the presence of breasts was quite noticeable in the X-rays. For this reason, we recommend future researchers add the sex of the patient, so the model can learn to avoid classifying breasts/nipples as diseases, but also so trained models can correlate diseases that more commonly plague members of a specific biological sex with a chest X-ray.

6 Conclusion/Future Work

Overall we've found that leveraging AI systems to supplement radiologists' diagnoses of many lung diseases with chest X-rays can be extremely meaningful especially considering the vast amounts of X-ray data radiologists have access to in terms of of X-rays, CT scans, and MRI images every day. We've found that drawing bounding boxes around likely areas of disease which would resemble a pre-filter system would both significantly reduce a radiologist's workload but also help them ignore negatives and hone in on positives, increasing their own diagnostic capacity.

We were able to train both a multi-label classifier (ResNet-18) and a diseased-area detector (Faster R-CNN + MobileNet backbone) on the NIH Kaggle dataset and were able to achieve a 95% val Acc_{IoU} with our classifier.

In future, if we had more time and team members we would love to have explore additional datasets expanding from our current Chest X-ray images. This could include CT scans and MRI images, expanding to other image based datasets abundant to radiologists.

7 Contributions

Sam: Object detection via training Faster-RCNN; wrote up the entire training and testing pipeline, as well as created visualizations. Attempted semi-supervised teacher-student model object detection training.

Ryan: Wrote up pipeline for training and testing classification models; created new metrics for multi-label classification and came up with confusion matrix visualization

Kiara: Dataset analysis (normalization constants, frequency analysis, label analysis), wrote up train/val dataset and visualization scripts; write-up lead and video lead

References

[1] A.J. DeGrave, J.D. Janizek, and SI Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nat Mach Intell*, 3:610–619, 2021.

[2] Ching Travers et al. Opportunities and obstacles for deep learning in biology and medicine. *Royal Society*, 15, 2018.

[3] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[4] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[5] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *CoRR*, abs/1801.09927, 2018.

[6] C.J. Kelly, A. Karthikesalingam, and M. et al. Suleyman. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*, 17, 2019.

[7] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, and D. F. Kallmes. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology*, 2015.

[8] A. Narin, C. Kaya, and Z. Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Anal Applic*, 24:1207–1220, 2021.

[9] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[10] M. Roberts, Thorpe Driggs, D., and M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nat Mach Intell*, 3:199–217, 2021.

[11] 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. *CoRR*, abs/1705.02315, 2017.

[12] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels, 2018.