# Detecting abnormalities on chest X-rays using deep neural networks

Swaroop Kumar M L

April 27, 2019

**Abstract**

Diseases like pneumonia and tuberculosis are leading causes of death worldwide. Although conclusive diagnosis requires other tests such as a sputum culture, chest radiography can be an important diagnostic aid and is routinely recommended since it is fast, affordable and highly sensitive. Moreover, automated detection of abnormalities on the chest X-ray can help in screening and severity-based prioritization. Due to the nature of the domain, it is also important that models not only make inferences but also generate explanations sufficient to convice a human expert.

Inspired by previous work, we develop algorithms that can detect abnormalities on the X-ray and explain these detections using weakly supervised localization. We establish baselines, benchmark against previous work and show that recent techniques such as mixup and progressive resizing can help to improve performance and generalization to other datasets. In terms of AUROC, we acheive performance competitive with previous work in a) detecting pneumonia-like and other abnormalities on the NIH chestX-ray14 dataset and b) detecting tuberculosis on the Shenzhen hospital dataset, and achieve state-of-the-art performance on the Montgomery county tuberculosis dataset. We also look for potential sources of bias and test our baselines with respect to gender, age and view position.

# Contents

# Chapter 1

# Introduction

## 1.1 Problem definition

Here I describe the problem and define it formally, including the need for model explainability.

## 1.2 Motivation

Here I talk about the social/economic aspects of these diseases, compare different diagnosis methods and how and why chest radiography, and our work in particular can help. Also, how this can be part of a bigger practical implementation integrated into existing radiologist workflow. More detail on the latter in section 9.3.

## 1.3 Previous work

A breif overview of previous work. A more detailed description will be in chapter 2, *Literature survey*.

## 1.4 Our work

Short overview of our experiments and results. A condensed version of chapters 5, 6 and 7

## 1.5 Report layout

Overall layout of the rest of the report.

# Chapter 2

# Literature survey

## 2.1 CheXNet and CheXNext

CheXNet and later, CheXNext from the Stanford ML group, are the most well-known works in this area. Our work is heavily derived from theirs.

## 2.2 Weakly supervised learning

I breifly describe the body of work around weakly supervised learning, and for our purposes, two forms of it:

1. Learning with inaccurate labels. For example, learning from labels which were generated algorithmically and which may therefore be inaccurate. Here, labels were extracted from radiology reports in natural language text.

2. Learning from imprecise labels. For example, learning to precisely localize objects or patterns with imprecise image-level labels. We use this to generate explanations.

## 2.3 Explainability

A short review of methods such as weakly supervised localization, LIME and SHAP which have been developed for generating explanations.

## 2.4 Fairness

Especially in deep learning, since a model is a black-box and can learn biases inherent in large datasets such as race and gender, a number of methods have been proposed to quantify and/or prevent this.

## 2.5 Learning at multiple scales

Several papers use multiple scales to improve performance.

## 2.6 Attention

The state-of-the-art for the NIH chestX-ray14 dataset uses a form of attention by training a seperate local branch on small patches of an image. I describe this and other methods that use attention.

## 2.7 Recurrent neural networls

A number of papers have shown that using recurrent neural networks to effectively make use of correlations between different abnormalities improves performance.

## 2.8 Generalizability

Atleast one other paper has studied how models in this domain generalize to other datasets. I describe their results.

## 2.9 Other methods

Other methods such as capsule networks and squeeze-and-excitation blocks have been used.

## 2.10 Tuberculosis

Here, I describe the previous work on tuberculosis detection, where the emphasis is on traditional techniques due to the small size of the dataset.

# Chapter 3

# Data

Here I describe all the datasets we use, as well as how we split them, and where and how we use k-fold cross validation

## 3.1 NIH CXR-14

The NIH chestX-ray 14 dataset, our primary dataset annotated with 14 different abnormalities including pneumonia.

### 3.1.1 Challenges and issues

Challenges with this dataset, and issues arising due to the fact that labels were extracted algorithmically from radiology reports in natural language.

## 3.2 Mendeley

This is the Mendelay pneumonia dataset of CXRs of children under 5 years, which we use to test generalization.

## 3.3 Szhenzhen and Montgomery

Szhenzhen and Montgomery county tuberculosis datasets. Shenzhen is used to train models and Montgomery is used to test generalization.

# Chapter 4

# Baselines

r

## 4.1 Model architecture

## 4.2 Training procedure

## 4.3 Weakly supervised localization

### 4.3.1 Saliency map to bounding box

### 4.3.2 LIME as an alternative

# Chapter 5

# Experiments

# Chapter 6

# Results

## 6.1 Comparison metrics

## 6.2 Comparison to previous work

## 6.3 Comparison to human radiologists

### 6.3.1 Caveats with comparison to human radiologists

## 6.4 Other important factors

Such as how well predicted probabilities correspond to actual severity, localization, and time

## 6.5 Examples

# Chapter 7

# Bias

## 7.1 Gender

## 7.2 Age

## 7.3 View position

# Chapter 8

# Conclusion

# Chapter 9

# Future work

## 9.1 Limitations of proposed work

## 9.2 Avenues for further research

## 9.3 Practical implementation and clinical relevance

## 9.4 More recent datasets

Such as CheXPert and PadChest