Detecting abnormalities on chest X-rays using deep neural networks

Swaroop Kumar M L
Department of Studies in Computer Science
University of Mysore

May 14, 2019

Abstract

Diseases like pneumonia and tuberculosis are leading causes of death world-wide. 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. The algorithm explains these detections by generating heatmaps pointing out areas of the image that most influenced it. 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. 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 look for potential sources of bias and test our baseline with respect to gender, age and view position.

Contents

Chapter 1

Introduction

1.1 Problem definition

The lungs are made up of small air-sacks called alveoli. When, for example, air in the alveoli is replaced with pus, blood and other fluids, referred to as consolidation and commonly caused by pneumonia, or when abscesses in the lung rupture forming cavities, indicating a tuberculosis infection, these are visible on the chest x-ray. See figure 1.1 for examples.

Radiologists are trained to look for signs of these abnormalities, use subtle visual features to differentiate among the various types, reason about their causes and help in diagnosis and treatment. An algorithm that can automatically detect these abnormalities can help by:

- Screening for patients with a particular disease in the general population
- Prioritizing patients for subsequent review by a trained radiologist
- Aiding a radiologist by being part of his/her workflow

However, building the hardware and software infrastructure for a clinically relevant system that is useful in practice is a problem which presents unique challenges of its own[1]. Our work focuses on the core algorithm. We divide the problem into, and explore, four sub-problems.

1.1.1 Classification

For our primary dataset, a large collection of chest x-rays annotated with 14 different abnormalities including pneumonia [1] (see section 3.1), we formulate

[1] We discuss possible implementations in section 9.3

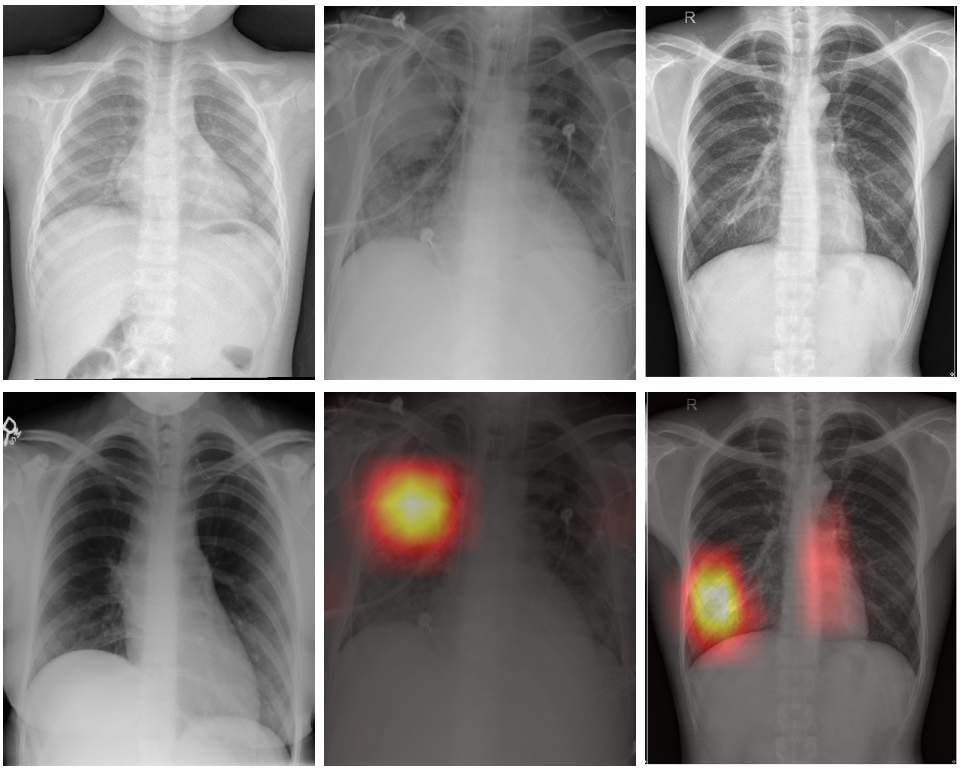

Figure 1.1: From left to right: The first column shows two *Normal* images. Columns 2 and 3 show images with *Pneumonia* and *Tuberculosis* respectively, the first row showing the original image and the second showing the same overlaid with heatmaps localizing the abnormalities, which we call *explanations*

the problem as a multi-class multi-label classification problem. Given the n-dimensional feature space $X = \mathbb{R}^n$, and a set of k class labels corresponding to k abnormalities $L = \{ l_1, l_2, l_3 \dots l_k \}$ the task is to learn a function $f : X \to 2^L$ from the training set $D = \{ (x_i, Y_i) \mid 1 \leq i \leq m \}$. For each example (x_i, Y_i), $x_i \in X$ is an n-dimensional feature vector $\{ x_{i1}, x_{i2}, x_{i3} \dots x_{in} \}$, each feature representing the intensity value of a single pixel of the input image, and $Y_i \subseteq L$ is the set of abnormalities associated with x_i. For an unseen instance x, the classifier $f(\cdot)$ predicts $f(x) \subseteq L$ as the set of abnormalities for x.

For the Shenzhen hospital tuberculosis dataset (see section 3.3), we formulate the problem as a binary classification problem. Again, suppose X is the n-dimensional feature space. The task is to learn a function $f : X \to \{ 0, 1 \}$ from the training set $D = \{ (x_i, y_i) \mid 1 \leq i \leq m \}$. For each $(x_i, y_i) \in D$, $x_i = \{ x_{i1}, x_{i2}, x_{i3} \dots x_{in} \}$ is an n-dimensional feature vector, each feature corresponding to the intensity value of a single pixel of the input image, and $y_i \in \{ 0, 1 \}$ is the corresponding label, 0 meaning *Normal* and 1 meaning *Tuberculosis*. Given an unseen $x \in X$, the classifier $f(\cdot)$ predicts $y \in \{ 0, 1 \}$ as being the label for x.

In both cases, $f(\cdot)$ is a deep convolutional neural network[2], specifically a variant of densenet[3] trained using the Adam[4] optimization algorithm (see sections 4.1 and 4.2 for discussions about the model architecture and training procedure). In the case of NIH ChestX-ray14, our primary dataset, the output of the network is a 14-dimensional vector $P = \{ p_1, p_2, p_3 \dots p_{14} \}$ where $p_k \in P$ corresponds to the k^{th} abnormality and $0 \leq p_k \leq 1$. A set of optimal thresholds $T = \{ t_1, t_2, t_3 \dots t_{14} \}$ is determined from the training set by maximizing the F1-score and applied to the output of the network so that the final output $Y \subseteq L$ is:

$$Y = \{ l_i \in L \mid p_i \in P \wedge t_i \in T \wedge p_i > t_i \} \tag{1.1}$$

In the case of the Shenzhen hospital tuberculosis dataset, the output of the network is a 2-dimensional vector $P = \{ p_1, p_2 \}$ where $0 \leq p_1 \leq 1$ and $0 \leq p_2 \leq 1$ and the final output y is:

$$y = \begin{cases} 0 & \text{if } p1 \geq p2 \text{ (Normal)} \\ 1 & \text{if } p2 > p1 \text{ (Tuberculosis)} \end{cases} \tag{1.2}$$

1.1.2 Explainability

Deep neural networks have outperformed previous methods in several domains. However, they remain black-boxes with millions of parameters, limiting their use in routine clinical practice. Several methods have been proposed to

1.1.3 Generalizability

1.1.4 Fairness

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

2.2 Weakly supervised learning

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

Methods to quantify learned bias along the lines of gender, race, etc.

2.5 Learning at multiple scales

Methods to effectively combine inferences from multiple scales.

2.6 Attention

Methods to allow models to selectively pay attention to parts of an image.

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

2.10 Tuberculosis

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

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

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

3.4 Montgomery

Chapter 4

Baselines

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

5.1 Data augmentation

5.2 Test-time augmentation

5.3 Higher resolution

5.4 Progressive resizing

5.5 Ensembling for scale-invariance

5.6 Ensembling saliency maps

5.7 Mixup

5.8 Self-training

5.9 Transfer learning for tuberculosis

5.10 Generalizability

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

Bibliography

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, *ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, 2017.

[2] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.