Detecting abnormalities on chest X-rays using deep neural networks

Swaroop Kumar M L

April 26, 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. There has been increasing interest in using deep learning for computer-aided diagnosis in both the machine learning and radiology communities. The community has also come to recognize that human experts sometimes disagree, which may lead to label noise, and that models should be required not only to make inferences but to also *explain* these inferences such that human experts can decide whether an inference can be trusted. There has been recent work on model explainability and methods to deal with label noise.

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 evaluate the effects of recent techniques such as mixup on performance and generalizability 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) in detecting tuberculosis on the Shenzhen hospital dataset, and we achieve state-of-the-art performance on the Montgomery county tuberculosis dataset. We also test our algorithms for bias with respect to gender, age and view position.

Contents

Chapter 1

Introduction

1.1 Problem definition

1.2 Motivation

1.3 Previous work

1.4 Our work

1.5 Report layout

Chapter 2

Literature survey

2.1 CheXNet and CheXNext

2.2 Weakly supervised learning

2.3 Explainability

2.4 Fairness

2.5 Learning at multiple scales

2.6 Attention

2.7 Generalizability

2.8 Other methods

2.9 Tuberculosis

Chapter 3

Data

3.1 NIH CXR-14

3.1.1 Challenges and issues

3.2 Mendeley

3.3 Szhenzhen and Montgomery

Chapter 4

Baselines

4.1 Model architecture

4.2 Training procedure

4.3 Unsupervised localization

4.3.1 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 Location-sensitive ensembling

5.7 Mixup

5.8 Label drift

Choose a better name

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

Scalability

9.4 More recent datasets