Detecting abnormalities on chest X-rays using deep neural networks

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

May 17, 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 active case finding, screening, and in cases where other tests are not available or are inconclusive. Due to the nature of the domain, it is also important that algorithms 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 a) transfer-learning from a large non-TB dataset dramatically improves TB detection, b) models in the domain show inferior performance on external data from a different hospital system but c) recent techniques such as mixup and progressive resizing improve performance and generalization. We achieve performance competitive with previous work in detecting pneumonia-like and other abnormalities on the NIH chestX-ray14 dataset and in detecting tuberculosis on the Shenzhen hospital dataset, and achieve state-of-the-art performance on the Montgomery county tuberculosis dataset. We evaluate our algorithms on bacterial and viral pneumonia separately, 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 be useful in numerous ways (see section 1.2).

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] We discuss possible implementations in section 9.3

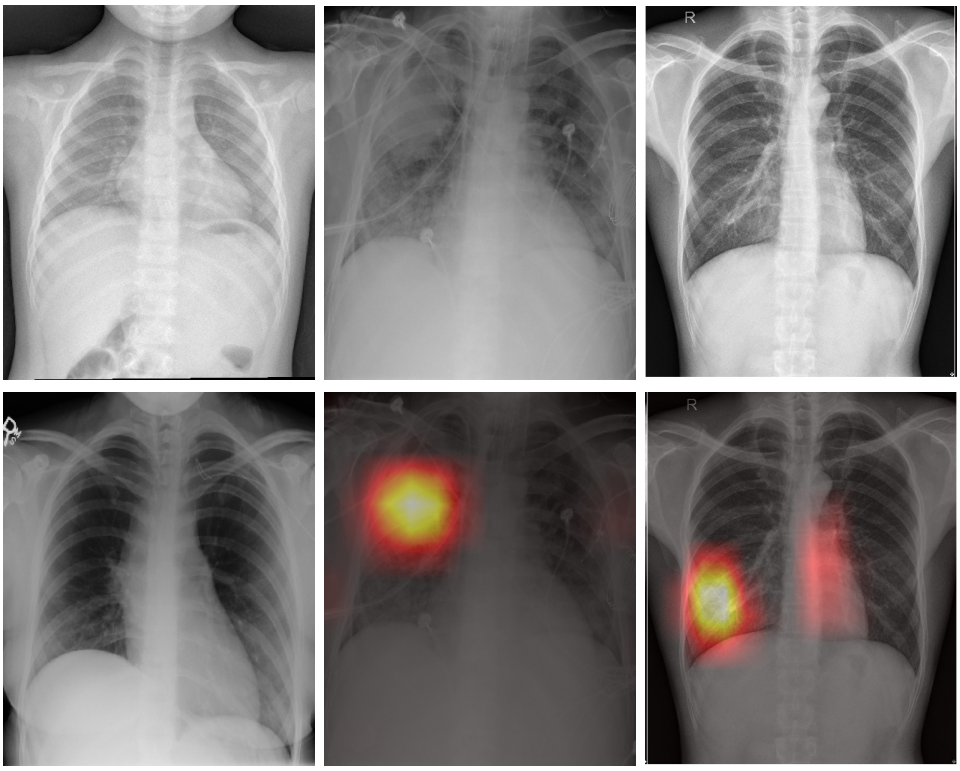

Figure 1.1: From left to right: The first column shows two *Normal* images. Columns 2 and 3 show images with *Pneumonia* and *Tuberculosis* respectively, the first row showing the original images and the second showing the same overlaid with heatmaps localizing the abnormalities, which we call *explanations*

1.1.1 Classification

For our primary dataset, a large collection of chest x-rays annotated with multiple abnormalities including pneumonia [1] (see section 3.1), we formulate the problem as a multi-class multi-label classification problem. Given the n-dimensional input feature space $X = \mathbb{R}^n$, and a set of c class labels corresponding to c abnormalities $L = \{ l_1, l_2, l_3 \dots l_c \}$ 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 $(x_i, Y_i) \in D$, $x_i \in X$ 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 input 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 $(x_i, y_i) \in D$, $x_i \in X$ is an n-dimensional feature vector $\{ x_{i1}, x_{i2}, x_{i3} \dots x_{in} \}$, 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 $f(x) \in \{ 0, 1 \}$ as being the label for x.

In both cases, $f(\cdot)$ consists of 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 by maximizing the network's F1-score on the training set 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, leading to a lack of trust and limiting their use in routine clinical practice. Several methods have been proposed to make these models more interpretable, broadly falling into two categories:

1. Methods that create a proxy model which behaves similarly to the original model, but is simpler and easier to understand. These include methods like LIME[5] and SHAP[6]. While these methods are model-independent, they tend to be very slow.

2. Methods that generate a saliency map which highlights a small portion of the input which is most relevant, in a single forward and backward pass through the network. These include methods like LRP[7], DeepLIFT[8], CAM[9] and Grad-CAM[10].

We use the CAM method and use the term *explanation* to mean a saliency map or heatmap. Given an input $x = \{ x_1, x_2, x_3 \dots x_n \} \in X$, and a set of class labels $L = \{ l_1, l_2, l_3 \dots l_c \}$, the task is to compute the attribution $A_j = \{ a_{j1}, a_{j2}, a_{j3} \dots a_{jn} \} \in \mathbb{R}^n$ for each class $l_j \in L$ where $a_{ji} \in A$ is a measure of the relevance of the i^{th} feature to the model's inference regarding the j^{th} class.

The network consists of a fully convolutional DenseNet backbone followed by an adaptive-average-pooling layer and a single fully-connected layer. The fully convolutional part of the network results in k w x h feature maps which are averaged along the width and height to form a k dimensional vector, which is fed to the single fully-connected layer with k input nodes and c output nodes. If f_i is the i^{th} feature map and w_i^j is the weight between the i^{th} input node and the j^{th} output node in the fully-connected layer, the saliency map for the j^{th} class M_j is

$$M_j = \sum_i w_i^j \, f_i \tag{1.3}$$

M_j is a w x h saliency map which is interpolated to the size of the input image and measures the relevance of each pixel to the model's decision regarding the j^{th} class, and can be visualized as a heatmap (see figure 1.1). This serves as an *explanation* of the model's inference, allowing physicians and radiologists to decide how much trust to invest in it.

1.1.3 Generalizability

A test set is considered representative of data that will be encountered in the external world and is used exclusively to evaluate a model. However, true generalization to new datasets may be lower than expected.

1. Since model design choices are based on previous work, methods in an application domain may overfit to one or a few popular datasets. However, [11] shows that this is not the case for CIFAR-10 despite years of methods being tested on this dataset

2. Two datasets may have different distributions. In the context of biomedical imaging, datasets may be collected from different hospital systems and machines. For example, in [12], Zech et al. show that models trained on data from one hospital system showed inferior performance on data from others.

3. The dataset used to train a model may have confounding variables that do not exist in other datasets. For example, in [12], Zech et al. also show that CNNs were able to directly detect the hospital system and department within a hospital system from a chest radiograph where saliency maps showed high activation in image corners. Since different departments and machines within a hospital system have different prevalence of a disease, the model may leverage these spurious correlations and fail to generalize

Since we observed the same phenomenon as in [12], of saliency maps showing high activation in image corners, we evaluated our models on external datasets from different hospital systems and studied how recent techniques affected generalization.

1.1.4 Fairness

Machine learning systems are increasingly being deployed in settings where they may inadvertently learn and leverage biases in the datasets, discriminate based on race, gender, etc. and amplify existing social inequities. For example, in [13], Bolukbasi et al. show that the popular word embedding space Word2Vec encodes gender bias. In [14], Buolamwini et al. show that facial recognition datasets are overwhelmingly composed of light-skinned individuals and that commercial gender classification systems performed worse for dark-skinned people and females, with a difference in accuracy of more than 30% between light-skinned males and dark-skinned females.

There has been substantial work in the research literature on fairness in ML on the development of statistical definitions of fairness [15]–[17] and algorithmic methods to measure and mitigate undesirable biases [17]–[19]. A simple measure against bias in various fields has been to hide variables like gender and race from a model, but complex machine learning models learn to use other correlated variables as proxy for hidden ones (for example, zipcode as correlated with race and the word 'women' in an institution's name as correlated with the gender of its students[20]). Moreover, hiding sensitive variables from researchers exacerbates the problem by limiting their ability to quantify and mitigate these biases. For example, in [21], Estava et al. found that convolutional neural networks are effective at detecting melanoma from images. However, without labels for skin characteristics such as color, accuracy of the model for different skin-types cannot be measured.

In this work, we measure the potential for discrimination by

1. Measuring the correlation of various abnormalities with gender and age group

2. Training models with architectures similar to the abnormality detection model, to identify gender and age group from images alone.

We test our baseline model's performance across genders and age groups.

1.2 Motivation

Pneumonia and tuberculosis are leading causes of death worldwide. According to the world health organization, pneumonia disproportionately affects children, accounting for 16% of all deaths of children under the age of 5 years. Tuberculosis is more prevalent in countries where many people live in absolute poverty with limited access to healthcare and in 2017 alone, caused 1.6 million preventable deaths.

The global End TB strategy aims for a 95% reduction in deaths due to TB by 2035 compared with 2015. Similarly, the National Strategic Plan (NSP) 2017-2025 sets out to achieve a rapid decline in deaths due to TB and emphasizes the importance of active case finding, that is, detection of TB cases early by seeking out people in targeted groups and scaling up cheap and high sensitivity TB diagnostic tests. The NSP has recommended three

tests: sputum smear microscopy, chest x-ray and the new CB-NAAT[2] test.

Conventionally, patients are screened for TB or pneumonia related symptoms, sputum examinations are recommended for those with positive symptoms, and chest x-rays are recommended for those who test negative in the sputum examination.

With automated detection, x-ray tests have the potential to be faster and significantly more affordable. They can be massively scaled up and used

1. For active case finding in high-risk populations, for example, with mobile x-ray vans[22]
2. As an initial screening test before or along with other tests such as a sputum examination
3. To aid a radiologist in her workflow by sorting her queue based on severity, suggesting areas to consider in an image, providing a second opinion, etc.

Our motivation in this regard is to further the goal of making automated abnormality detection systems such as ours clinically relevant by making them more accurate and explainable, testing their ability to generalize to other hospital systems and making sure they do not discriminate based on gender, age group, etc.

1.3 Previous work

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.

[2]CB-NAAT or Cartridge Based Nucleic Acid Amplification Test is a molecular test and is known as GeneXpert outside India

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.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 2016, pp. 1135–1144.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, e0130140, 2015.

[8] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 3145–3153.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[11] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?" *arXiv preprint arXiv:1806.00451*, 2018.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS medicine*, vol. 15, no. 11, e1002683, 2018.

[13] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[14] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[15] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[16] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, ACM, 2012, pp. 214–226.

[17] M. Hardt, E. Price, N. Srebro, *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[18] A. Agarwal, A. Beygelzimer, M. Dudık, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

[19] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[20] J. Dastin, *Insight - amazon scraps secret ai recruiting tool that showed bias...* Oct. 2018. [Online]. Available: `https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH`.

[21] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[22] A. Modi and R. Suresh, *Scaling up tb screening with ai: Deploying automated x-ray screening in remote regions*, Apr. 2019. [Online]. Available: `http://blog.qure.ai/notes/scaling-up-tb-screening-with-ai`.