Detecting abnormalities on chest X-rays using deep neural networks

Swaroop Kumar M L

Abstract

Diseases like pneumonia and tuberculosis are leading causes of death world-wide and chest radiography can be an important diagnostic aid since it is fast, affordable and highly sensitive. Moreover, automated detection of abnormalities on the chest X-ray can aid the radiologist by being part of a larger system integrated into her workflow, and help in triaging and active case finding. Inspired by previous work, we develop algorithms that can detect abnormalities on the x-ray and explain 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, c) mixup, a recently proposed data augmentation technique, and progressive resizing, improve performance and generalization, d) models trained to detect tuberculosis, especially those pre-trained on a large non-TB dataset, tend to overdiagnose TB, and e) saliency maps for most abnormalities show high activations at image corners around regions of the image containing shadows of metal tokens. 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. We compare performance with and without lung-segmentation, look for potential sources of bias by training networks to identify gender, age and view-position from images alone, test our baseline for variable performance with respect to these, and evaluate our models on viral and bacterial pneumonia separately.

1 Introduction

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[1]. Tuberculosis is more prevalent in countries where many people live in absolute poverty[2] with limited access to healthcare and in 2017 alone, caused 1.6 million preventable deaths[3].

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[1] 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[4]
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.

2 Previous work

In [5], Wang et al. collect chest x-ray images and their associated reports from the PAC system of the National Institutes of Health and mine labels from the reports algorithmically. Previous work has explored methods to improve classification performance. In [6]–[9], the authors use attention-guided

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

learning to allow a network to concentrate on abnormal regions of the image. [7] also uses curriculum learning and presents images in increasing order of difficulty. However, [10] uses attention to hide the most salient regions, allowing the network to pay attention to other areas. [8], [11] seek to exploit correlations between abnormalities, [11] by using an LSTM and [8] by extracting saliency maps at an intermediate layer and providing these as input to subsequent layers.

There has also been work on improving localization by combining feature maps from multiple layers of the network [12], [13]. [13] learns a set of *layer relevance weights* for each class, and [12] applies a DenseNet per resolution orthogonal to a standard ResNet followed by upsampling and concatenation.

In [14], Rajpurkar et al. train a variant of DenseNet on the NIH chestX-ray14 dataset relabeled using an ensemble of classifiers and report super-human performance for several abnormalities, comparing board-certified radiologists and the algorithm on a test set labeled by consensus of three cardiothoracic subspeciality radiologists.

For TB detection, previous work such as [15]–[17] have explored various feature extraction techniques, feature selection strategies and classifiers such as logistic regression and SVM. [18]–[21] train deep convolutional neural networks and ensemble these. These methods have also explored the usefulness of segmentation of chest regions. Due to the lack of large publicly available datasets, work in this domain, especially the application of deep learning methods, has been limited. Moreover, results are less relevant to clinical practice as models trained on small two-class datasets are prone to over-diagnose.

3 Datasets

We use the NIH ChestX-ray14 dataset[5] and the Shenzhen hospital tuberculosis dataset[22] to train models to detect pneumonia and other abnormalitites, and tuberculosis respectively. We then use two external datasets, the Guangzhou medical center pediatric pneumonia dataset[23] and the Montgomery county tuberculosis dataset[22] as *external* datasets to test the ability of these models to generalize to other hospital systems.

We split the NIH CXR-14 into train, validation and test sets roughly in the ratio 70:10:20. We make sure that there is no patient overlap, that is, all images of a patient are in the same subset since patient overlap may lead to overfitting. We use the standard test set of the Guangzhou dataset and split the rest of the dataset into train and validation sets in the ratio 80:20. However, when using this as an external dataset, we use the entire dataset. Considering the small size of the Shenzhen and Montgomery datasets, we create 9 folds and report average and standard deviation of metrics. Each fold contains all the images split into train, validation and test sets in the ratio 70:10:20. However, when using the Montgomery dataset as an external dataset, we ignore these folds and test on the entire dataset.

4 Model

We replace the final fully connected layer of 121-layer dense convolutional neural network with one that has either 14 outputs (for the NIH CXR-14 dataset) or 2 outputs (for all other datasets) after which we apply a sigmoid non-linearity. The fully convolutional backbone of the network results in k w x h feature maps and is followed by a global-average-pooling layer where the k feature maps are averaged along the width and height to form a k dimensional vector. This makes the network independent of input image size and allows us to use the progressive-resizing method. The network's connectivity pattern improves the flow of information and gradients and has fewer parameters, making it possible to train very deep networks. The architecture of the model, specifically the fact that the fully convolutional part of the network is followed by a single fully connected layer, forces the model to learn to localize abnormalities given only weak labels (presence or absence of an abnormality).

5 Training and inference

We use the Adam optimization algorithm and start the training with an initial learning rate a factor of 10 smaller than the learning rate at which the training loss begins to increase, when the learning rate is increased linearly (using a learning rate finder). We divide the learning rate by 10 if the validation loss plateaus (does not decrease over 5 iterations), and stop training when the validation loss has stopped decreasing. When using k-fold cross validation, we train k different networks on each of the k folds and report average and standard deviation of performance.

At the multi-class multi-label classification task (for the NIH CXR-14 dataset), we merge the train and validation sets after training and use it to compute optimal thresholds for each abnormality by optimizing for the class-specific F1-score. Although it is possible to compute optimal thresholds for the

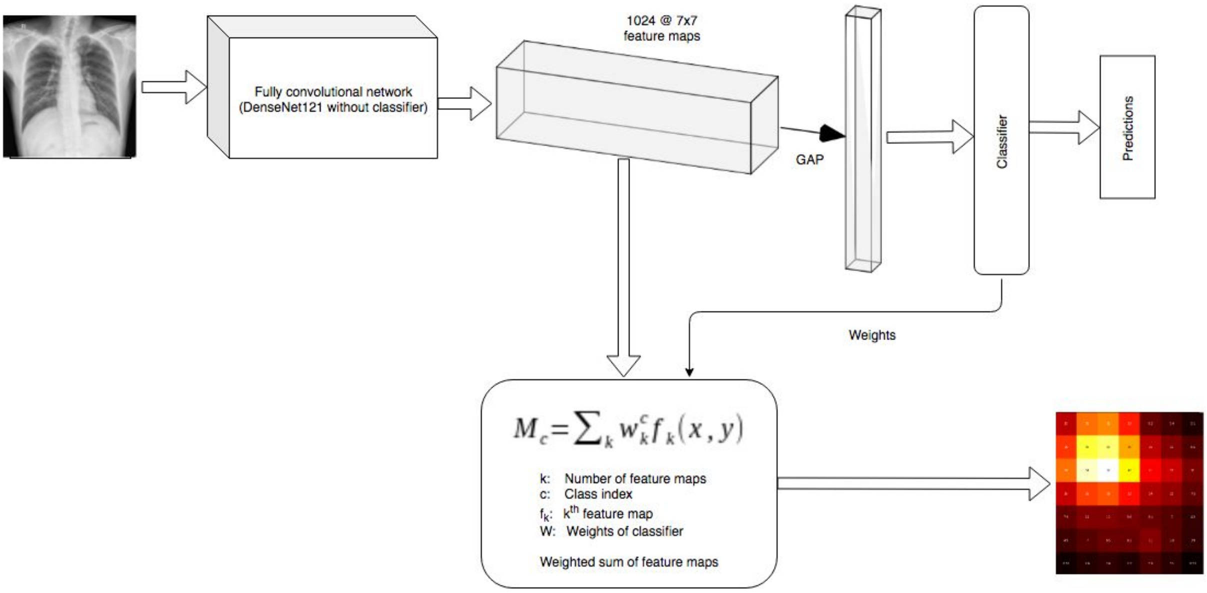
Fully convolutional network
(DenseNet121 without classifier)
1024 @ 7x7
feature maps
GAP
Classifier
Predictions
Weights
$M_c=\sum_k w_k^c f_k(x,y)$
k: Number of feature maps
c: Class index
f_k: k^{th} feature map
W: Weights of classifier
Weighted sum of feature maps

Figure 1: Basic architecture of the model

classification task since we have ground truth labels, it is not possible to do so for the localization task since we only have weak labels and not precise locations. At binary classification tasks, since the network has two output nodes, we do not compute thresholds but simply consider as the proper output the class whose corresponding output node has higher activation.

6 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: a) methods that create a proxy model which behaves similarly to the original model, but is simpler and easier to understand, which include methods like LIME[24] and SHAP[25]which are model-independent but tend to be very slow. and b) 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 networkwhich include methods like LRP[26], DeepLIFT[27], CAM[28] and Grad-CAM[29].

We use the CAM method. To compute explanations, we save the feature maps resulting from the final convolutional layer during a forward pass and perform a weighted sum of these feature maps using the weights of the final fully-connected layer between each of the feature maps and the desired output node, as follows.

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

$$M_j = \sum_i w_i^j f_i \tag{1}$$

M_j is a w x h saliency map which we interpolate to the size of the input image and visualize as a heatmap.

We use a region-growing algorithm to determine bounding boxes given a saliency map. Specifically, we first threshold the saliency map and using the maximum element as a seed point, grow a region around it, including all non-zero neighbours, and repeat the same until all non-zero elements are included in a region, each time choosing as seed the maximum element not included a region. For each region, we determine a bounding box as the smallest rectangle which encloses the entire region (for example, see 2).

7 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. Two datasets may have different

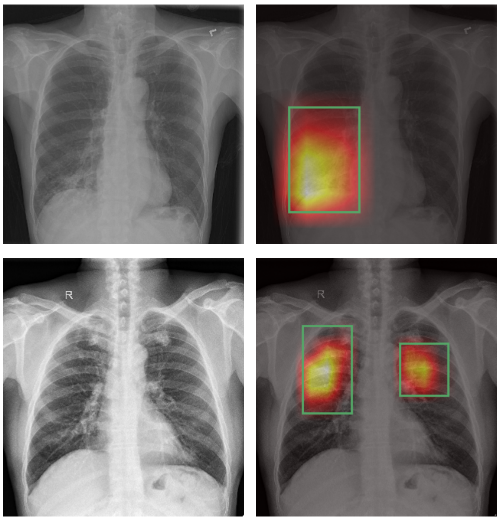

Figure 2: Examples of saliency maps with corresponding bounding boxes drawn. The first column shows original x-ray images and the second row shows saliency map and bounding boxes overlaid on the image. Row 1: Pneumonia. Row 2: Tuberculosis.

distributions. In the context of biomedical imaging, datasets may be collected from different hospital systems and machines. For example, in [30], Zech et al. show that models trained on data from one hospital system showed inferior performance on data from others. The dataset used to train a model may have confounding variables that do not exist in other datasets. For example, in [30], 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. We observed the same phenomenon in networks that were trained to detect abnormalities (figure 3 shows average saliency maps or explanations weighted by predicted probability). 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.

We evaluate the ability of our baselines to generalize to other hospital systems. At the problem of pneumonia detection, we train on the NIH CXR-14 dataset and test on both the internal test set and the external Guangzhou dataset, a pediatric pneumonia dataset from a different hospital system. At the problem of tuberculosis detection, we train on the Shenzhen tuberculosis and and test on both the internal test set and the external Montgomery tuberculosis dataset from a different hospital system.

We find that the baseline model trained to detect pneumonia on the NIH CXR-14 dataset performs better on the Guangzhou dataset than the internal test set but worse that a model trained exclusively on the external dataset, and the baseline model trained to detect tuberculosis on the Shenzhen dataset shows inferior performace on the external dataset, but which is better than a model trained exclusively on the external dataset.

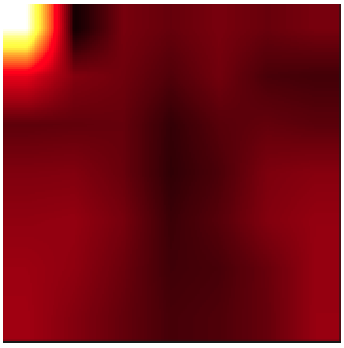

Figure 3: Average saliency maps for *pneumonia* of the baseline model trained on the NIH CXR-14 dataset.

8 Transfer-learning

We initialize the weights of a network with those of a similar network trained on a large dataset of millions of natural images, ImageNet and compare this with random initialization (we use the Kaiming He initialization method[31]) and show that pretraining on ImageNet significantly improves performance on both internal and external datasets.

On the Shenzhen hospital tuberculosis dataset, we replace the weights of a network with those of a similar network a) trained on ImageNet (a large non-x-ray dataset of natural images) and b) trained on ImageNet and then on NIH CXR-14 (a smaller non-tb x-ray dataset), and observe that pretraining on the NIH CXR-14 dataset both improves performance on both the internal test set and helps to close the generalization gap.

9 Mixup

Mixup, a recently proposed data augmentation technique [32], has been shown to be effective at regularizing models and combating label noise. Instead of using raw images, we feed the model a linear combination of two images not necessarily from the same class. If I_1 and I_2 are two images, we feed the network a linear combination $M = t \cdot I_1 + (1-t) \cdot I_2$ where t is drawn from a beta distribution parameterized by some α. The expected output for M is $t \cdot y_1 + (1-t) \cdot y_2$ where y_1 and y_2 are the targets for I_1 and I_2 respectively. Mixup improved performance on both the NIH CXR-14 dataset and the Shenzhen tuberculosis dataset. It also improved generalization to external datasets.

10 Progressive resizing

For the NIH CXR-14 dataset, we first train on 224 x 224 images until the validation loss plateaus. We then re-train the same network on 256 x 256 images, 288 x 288 images, etc. upto 512 x 512 with a smaller learning rate and for fewer epochs. For the Shenzhen hospital tuberculosis dataset, we first train on 224 x 224 images, until the validation loss plateaus, and retrain the same network on 448 x 448 images and then on 672 x 672 images.

The intuition behind progressive resizing is that first training on lower resolution images is equivalent to pre-training and is better than training on high resolution images from scratch. It may also make networks more robust to scale variation and behave similar to data augmentation preventing overfitting. We observe that progressive resizing consistently improves performance and generalization to external datasets.

11 Fairness

We look for potential sources of bias such as gender, age and view-position by a) looking at the distribution of ground truth by gender, age and view-position and variable rates of abnormalities, b) training models with architecture similar to our baseline model to predict these from images alone. We then measure variable performance of our baseline model across gender, age-group and view-position.

11.1 Age

Age follows a rougly gaussian distribution with a mean of 46.17 years and standard deviation of 16.73. On dividing patients into 10 age groups (0-9 years, 10-19 years . . . 90-99 years), abnormality rate increases with age, with a maximum rate of 57.4% for the age group 80-89 years and a minimum rate of 38.7% for the age group 0-9 years. *No finding* is negatively correlated with age, with a PMCC of -0.07. When broken down into 3 age groups (less than 25 years, between 25 and 65 years, and greater than 65 years) and by specific abnormalities, *Hernia* is 2.8 times more likely if the patient is old-aged (more than 65 years old) and *Pneumonia* is 1.4 times more likely if the patient is young (less than 25 years old).

However, a network with a similar architecture with 3 output nodes trained to detect the age group from x-ray images predicted the most common 2^{nd} age group (25 to 65 years) for every image. A network with a similar architecture with a single output node trained to predict age as a continuous variable achieved a mean absolute error of 10.9 years, which is not significantly better than the mean absolute deviation of a gaussian distribution with the same mean and standard deviation as that of patient-age, 13.3, meaning that the model's predictions are not much better than a naive algorithm which predicts the mean age of 46 years for every image.

We evaluated our baseline for variable performance for each age group and found that the model showed similar performance (in terms of AUROC) for each.

11.2 Gender

The male to female ratio is approximately 1.28, with similar rates of abnormality, 53.5% and 54.1% for males and females respectively, and *No finding* is only weakly correlated with *female*, with a PMCC of 0.006. However, when broken down by specific abnormalities, the ratio of the posterior probability of an abnormality given the gender to its prior probability shows significant variation, with *Hernia* becoming 1.3 times more likely and *Cardiomegaly* becoming 1.2 times more likely if the patient is female (pregnancy is a common cause of *Cardiomegaly*).

A similar network (with the same architecture, the only difference being 2 output nodes in the final layer instead of the 14) trained to identify gender from x-ray images on this dataset acheived an accuracy of 93.8% (AUROC of 98.9%) on this task when trained for a single epoch. Saliency maps showed high activations at and around regions of the image containing female breasts.

Although this does not necessarily mean that our abnormality-detection models are biased, the two findings above show that some bias exists in the dataset and that these models are capable of exploiting these. We evaluated our baseline model for variable performance for males and females and found that the model showed similar performance (in terms of AUROC) for both genders.

11.3 View-position

The PA (posterioanterior) view is preferred over the AP (anterioposterior) view. However, the AP view is usually chosen over the PA view for younger children and is necessitated for very ill patients who cannot stand erect. The PA view is more common in the NIH CXR-14 dataset with 60% of the images showing the PA view. Abnormality rate is higher for the AP view (52.8%) compared to the PA view (41.6%), and *No finding* is positively correlated with PA with a PMCC of 0.11. When broken down by specific abnormalities, the ratio of the posterior probability of an abnormality given the view-position to its prior probability shows significant variation, with *Edema* and *Consolidation* becoming 2.2 times and 1.7 times more likely respectively if the x-ray image shows AP view.

Moreover, a network with a similar architecture with 2 output nodes trained to identify the view-position from x-ray images acheived an accuracy of 98.7% (AUROC of 99.7%) on this task when trained for a single epoch. Saliency maps showed high activations at and around the anterior aspect of the ribs and around shadows of tokens on x-ray that identified the machine as being a *portable* machine.

We evaluated our baseline for variable performace for each view-position and found that the model showed similar performance (in terms of AUROC, sensitivity and specificity).

12 Over-diagnosis

The rate of of tuberculosis in the NIH CXR-14 dataset is unknown since tuberculosis is not one of the labels. However, assuming a baseline rate of less than 1%, we observe that models trained to detect tuberculosis on the Shenzhen hospital tuberculosis dataset tend to over-diagnose when tested on the NIH CXR-14 dataset. The rate of over-diagnosis is especially high when models are pre-trained on the NIH CXR-14 dataset.

13 Viral vs bacterial pneumonia

Images in the Guangzhou pediatric pneumonia dataset that show manifestations of pneumonia are further categorized as *Viral* and *bacterial*. Since viral and bacterial pneumonia present different levels of emergency and warrant different courses of treatment, we evaluate our models for variable performance on these categories to ensure that they are not biased toward one or the other type. We found that the models trained on the NIH CXR-14 dataset were better at detecting viral pneumonia than bacterial pneumonia.

14 Segmentation and centering

The Montgomery dataset includes hand-annotated segmentation masks of both the left and right lungs for each image. We evaluate models trained on the Shenzhen hospital tuberculosis dataset, on the Montgmomery dataset after a) segmenting the lung regions and b) segmeting the lung regions and cropping to the smallest rectangle which encloses both the lungs. Across multiple models and trials and averaged across 9 folds, we failed to see a significant increase or decrease in performance among models trained on un-segmented images, segmented images and segmented and cropped images.

15 Results

For the NIH CXR-14 dataset, we achieve performance competitive with previous work and show improvements over our baseline. See table 1 for the results. Rajpurkar et al. in [33] measured human performance in terms of AUROC for each disease, using the majority vote of 3 independent board-certified cardiothoracic specialist radiologists (average experience 15 years) as ground truth, and measure the the performance of 6 BC radiologists from 3 academic institutions (average experience 12 years) and 3 senior radiology residents by fitting a curve to these 9 radiologists' operating points and calculating the area under it. We compare our models with human radiologist performance and find that the model's performance is on average within 2% of that of human radiologists (see table 2 for the comparison).

On the Shenzhen and Montgomery datasets, we achieve performance comparable to previous work and show improvement over our baseline. See tables 3 and 4 for the results.

Authors	**AUROC**
Wang et al. (2017)	0.738
Y. Shen et al.	0.775
H. Wang et al. (ChestNet)	0.781
P. Kumar et al.	0.792
Yao et al. (2017)	0.803
Y. Tang et al.	0.805
S. Guendel et al.	0.807
Yan et al.	0.83
X. Xu et al. (DeepCXray)	0.832
Rajpurkar et al. (CheXNet)	0.841
B. Zhou et al.	0.842
Rajpurkar et al. (ChexNext)	0.849
Our model	**0.856**
Q. Guan et al.	0.871

Table 1: Comparison to previous work on the NIH CXR-14 dataset

Abnormality	**AUROC**			
	Baseline	Ensemble	Radiologist	Difference (%)
Atelectasis	0.823	0.839	0.808	-3.06
Cardiomegaly	0.899	0.916	0.888	-2.79
Effusion	0.881	0.89	0.9	0.96
Infiltration	0.705	0.72	0.734	1.39
Mass	0.857	0.868	0.886	1.76
Nodule	0.779	0.817	0.899	8.17
Pneumonia	0.767	0.765	0.823	5.83
Pneumothorax	0.881	0.895	0.94	4.46
Consolidation	0.822	0.819	0.841	2.22
Edema	0.911	0.902	0.91	0.83
Emphysema	0.913	0.944	0.911	-3.33
Fibrosis	0.824	0.854	0.897	4.31
Pleural Thickening	0.81	0.805	0.779	-2.59
Hernia	0.906	0.944	0.985	4.1
Average	**0.841**	**0.856**	**0.8715**	**1.59**

Table 2: Comparison to human radiologists on the NIH CXR-14 dataset

Authors	**AUROC**	**Accuracy**
Jaeger et al	0.9	0.841
Hwang et al	0.93	0.837
Lopez and Valiati	0.926	0.846
MT Islam et al	0.94	0.9
Haloi et al	0.949	
Liu et al (ResNet-152)	0.967	0.923
Liu et al (Inception-ResNet-v2)	0.983	0.917
Vajda et al	0.99	0.957
Our baseline	**0.956**	**0.902**
Our best model **Pretrained on NIH CXR-14** **with mixup** $\alpha = 0.4$	**0.985**	**0.949**

Table 3: Comparison to previous work on the Shenzhen tuberculosis dataset

Authors	**AUROC**	**Accuracy**
Jaeger et al	0.869	0.783
Lopez and Valiati	0.926	0.826
Liu et al (Inception-ResNet-v2)	0.957	0.844
Liu et al (ResNet-152)	0.951	0.890
Vajda et al	0.870	0.783
Our baseline	**0.871**	**0.755**
Our best model **Pre-trained on NIH CXR-14** **(480 x 480)**	**0.957**	**0.89**

Table 4: Comparison to previous work on the Montgomery tuberculosis dataset

16 Conclusion

Deep learning has surpassed human performance in several domains, especially in computer vision, and has recently been successfully applied to various tasks in the medical imaging domain, particularly in radiology. We develop algorithms that can detect abnormalities on the chest x-ray and explain these detections by generating saliency maps using the CAM method. We establish baselines and show that recent techniques improve performance and acheive performance competitive with previous work and human radiologists as measured by Rajpurkar et. al.

However, we also show that a) models in the domain show inferior performance on external data from a different hospital system, i.e, fail to generalize, b) models trained to detect tuberculosis, especially those pre-trained on a large non-TB dataset, tend to overdiagnose TB, and c) saliency maps for most abnormalities show high activations at image corners around regions of the image containing shadows of metal tokens. Further, deep neural networks were able to identify gender and view-position accurately from images alone and may learn to unfairly exploit correlations between these and specific abnormalities. Therefore, we call for further research to explore these and other challenges, improve model-interpretability and generalization to other hospital systems.

References

[1] *Pneumonia*, Nov. 2016. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia.

[2] *Tb and poverty.* [Online]. Available: https://www.tbalert.org/about-tb/global-tb-challenges/tb-poverty/.

[3] *Tuberculosis (tb)*, Sep. 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis.

[4] 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.

[5] 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," May 2017. DOI: 10.1109/CVPR.2017.369. arXiv: 1705.02315. [Online]. Available: http://arxiv.org/abs/1705.02315%20http://dx.doi.org/10.1109/CVPR.2017.369.

[6] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification," Jan. 2018. arXiv: 1801.09927. [Online]. Available: http://arxiv.org/abs/1801.09927.

[7] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs," in, Springer, Cham, Sep. 2018, pp. 249–258. DOI: 10.1007/978-3-030-00919-9_29. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00919-9%7B%5C_%7D29.

[8] H. Wang and Y. Xia, "ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography," Jul. 2018. arXiv: 1807.03058. [Online]. Available: http://arxiv.org/abs/1807.03058.

[9] E. Pesce, P.-P. Ypsilantis, S. Withey, R. Bakewell, V. Goh, and G. Montana, "Learning to detect chest radiographs containing lung nodules using visual attention networks," Dec. 2017. DOI: 10.1016/j.media.2018.12.007. arXiv: 1712.00996. [Online]. Available: http://arxiv.org/abs/1712.00996%20http://dx.doi.org/10.1016/j.media.2018.12.007.

[10] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays," in, Springer, Cham, Sep. 2018, pp. 589–598. DOI: 10.1007/978-3-030-00934-2_66. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00934-2%7B%5C_%7D66.

[11] L. Yao, E. Poblenz, D. Dagunts, arXiv preprint arXiv ..., and undefined 2017, "Learning to diagnose from scratch by exploiting dependencies among labels," *arxiv.org*, [Online]. Available: https://arxiv.org/abs/1710.10501.

[12] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions," Mar. 2018. arXiv: 1803.07703. [Online]. Available: http://arxiv.org/abs/1803.07703.

[13] S. Sedai, D. Mahapatra, Z. Ge, R. Chakravorty, and R. Garnavi, "Deep Multiscale Convolutional Feature Learning for Weakly Supervised Localization of Chest Pathologies in X-ray Images," in, Springer, Cham, Sep. 2018, pp. 267–275. DOI: 10.1007/978-3-030-00919-9_31. [Online]. Available: http://

link.springer.com/10.1007/978-3-030-00919-9%7B%5C_%7D31.

[14] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, no. 11, A. Sheikh, Ed., e1002686, Nov. 2018, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002686. [Online]. Available: http://dx.plos.org/10.1371/journal.pmed.1002686.

[15] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Zhiyun Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and C. J. McDonald, "Automatic Tuberculosis Screening Using Chest Radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, Feb. 2014, ISSN: 0278-0062. DOI: 10.1109/TMI.2013.2284099. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24108713%20http://ieeexplore.ieee.org/document/6616679/.

[16] U. Lopes and J. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in Biology and Medicine*, vol. 89, pp. 135–143, Oct. 2017, ISSN: 0010-4825. DOI: 10.1016/J.COMPBIOMED.2017.08.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482517302548.

[17] S. Vajda, A. Karargyris, S. Jaeger, K. Santosh, S. Candemir, Z. Xue, S. Antani, and G. Thoma, "Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs," *Journal of Medical Systems*, vol. 42, no. 8, p. 146, Aug. 2018, ISSN: 0148-5598. DOI: 10.1007/s10916-018-0991-9. [Online]. Available: http://link.springer.com/10.1007/s10916-018-0991-9.

[18] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," G. D. Tourassi and S. G. Armato, Eds., vol. 9785, International Society for Optics and Photonics, Mar. 2016, 97852W. DOI: 10.1117/12.2216198. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2216198.

[19] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks," May 2017. arXiv: 1705.09850. [Online]. Available: http://arxiv.org/abs/1705.09850.

[20] M. Haloi, K. R. Rajalakshmi, and P. Walia, "Towards Radiologist-Level Accurate Deep Learning System for Pulmonary Screening," Jun. 2018. arXiv: 1807.03120. [Online]. Available: http://arxiv.org/abs/1807.03120.

[21] J. Liu, Y. Liu, C. Wang, A. Li, B. Meng, X. Chai, and P. Zuo, "An Original Neural Network for Pulmonary Tuberculosis Diagnosis in Radiographs," in, Springer, Cham, Oct. 2018, pp. 158–166. DOI: 10.1007/978-3-030-01421-6_16. [Online]. Available: http://link.springer.com/10.1007/978-3-030-01421-6%7B%5C_%7D16.

[22] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[23] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” Oct. 2017. arXiv: `1710.09412`. [Online]. Available: `http://arxiv.org/abs/1710.09412`.

[33] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,” *PLoS Medicine*, vol. 15, no. 11, A. Sheikh, Ed., e1002686, Nov. 2018, ISSN: 15491676. DOI: `10.1371/journal.pmed.1002686`. [Online]. Available: `http://dx.plos.org/10.1371/journal.pmed.1002686`.