CS230

Exploring the Efficacy of Using a Neural Network Trained on Non-Tuberculosis Chest X-Rays for Detecting Tuberculosis

Eugene Tian
Stanford University
Department of Computer Science
etian511@.stanford.edu

Pablo Ocampo
Stanford University
Department of Computer Science
ocampo@stanford.edu

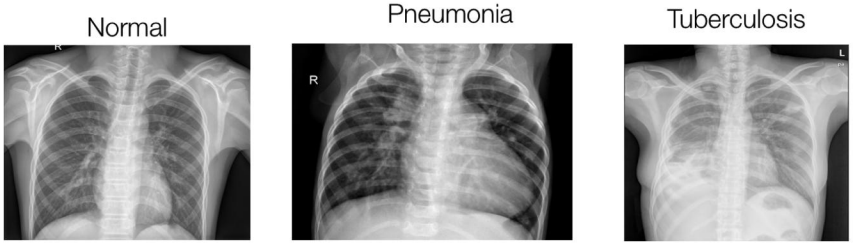
Abstract

Traditionally in health care, convolutional neural networks, or CNNs for short, are trained on normal images and images of a disease so that the model can detect the presence of that disease in an unlabeled image. However, reliable images of a specific disease may not always be available for training, so there exists a need for CNNs to be able to detect abnormalities that may not be present when training. This project explores the limitations of CNNs by testing how well a CNN trained on pneumonia images performs on tuberculosis detection. Afterwards, we trained our CNN on images of a variety of diseases not including tuberculosis and tested it on images of tuberculosis, and compared our two accuracies. We found that our CNN trained on a variety of diseases performed better at detecting tuberculosis than our CNN trained just on pneumonia.

1 Introduction

With deep learning and artificial intelligence becoming more and more prominent in modern technology, health care and biomedical data is coming to a new era as well. Medical imaging is a particularly expensive and time intensive section of healthcare, therefore Convolutional Neural Networks in particular have been at the center of attention as it takes advantage of big biomedical data to detect diseases efficiently and cost effectively. Our project focuses on lung diseases. Millions of people in the United States have a lung disease, and if all lung diseases are added together, it is the third biggest killer in the United States. With our project, we hope to explore how CNNs can detect the presence of Tuberculosis. Additionally, traditionally CNNs are trained on normal images and images of a disease so that when an unlabeled image is presented the CNN can detect whether the image has the disease or not. However, sometimes there may not be enough reliable images of a disease for a CNN to train on. Therefore, there is a need for CNNs to be able to effectively detect abnormalities when the training set does not necessarily include a specific disease. This project attempts to explore the limitations of CNNs by testing how well neural networks trained on images of pneumonia and images of normal lungs performs on images of tuberculosis. The input to our CNN was an unlabeled image, and the output was whether or not that image had pneumonia. Additionally, we trained a CNN with images of a variety of diseases and tested our performance on images of tuberculosis in order to explore how well CNNs can learn the difference between a normal lung and an abnormal lung. The input to this CNN was an

unlabeled image, and the output was whether the lungs were abnormal or not. The following is an example of the sorts of distinctions the model would have to make.



Furthermore, we are both in CS 221 right now and are sharing code infrastructure for lung classification between our two projects. The part of this project that was shared with the 221 project was the network that we trained using the smaller data set of pneumonia images. This network took a relatively short time to train and was given the simple task of binary classification between pneumonia lungs and healthy lungs, for which we didn't need to do too much fine tuning. The reason why we feel it's a good idea to include the results of that in this project is because we see this project as an extension and entirely different approach to the problem for which we wish to compare success rate; that is, using a more robust network that does not perform very well on classification of any particular disease but rather generalizes slightly better when detecting a more generic and simply unhealthy lung.

2 Related Work

The problem of lung disease detection using CNNs is well researched and documented. In [8], the study used a non convolutional neural network. While the model did demonstrate the possibility to convert radiologists' viewing into computer algorithms, it has weaknesses in that background information was not adequately reduced or learned with the neural network. Additionally, the neural network was not made to mimic the vision type network. It was clever as it used preprocessing to define suspected nodule areas. However, the use of non convolutional network did not adequately reduce background noise and caused accuracy to suffer.

[7] instead used CNNs for lung nodule detection. In their approach, like [8] they cleverly preprocessed images in order to enhance the presence of a nodule if there is one. This was done in a double-matching method where the first matching process enhances the detectability; furthermore there was a second matching procedure that evaluates the roundness of the suspected area in order to reject the artifact nodule if there is over-enhancement. This is a clever method as it makes learning for the CNN easier as the abnormality is more clearly pronounced in the images. Additionally, it was also smart to include a step to reject over-enhancement, in the case that preprocessing caused nodules to be more pronounced when they should not have been. This is a step that most preprocessing does not take into account. [10], like [7], uses a CNN, but instead it uses a 3-D CNN trained on weakly labeled data to detect lung nodules. This was effective as the 3D lung segmentation could eliminate the air tracts which are a primary cause of false positives. However, it's weakness is that it does not use 3D transforms of existing labels, which has been done for 2D CNNs. [11], also uses a CNN, but it uses pre-trained CNNs for tuberculosis detection. While their pre-trained networks did not require an expensive and time-consuming training set or need a large dataset, it's performance was inferior to a network trained specifically for the desired tasks. Additionally, it did not use fine-tuned weights, which also caused accuracy to suffer.

Unlike [7] and [8], [9] utilizes CT scans as inputs to a model to detect patterns for interstitial lung diseases. Like [7] [10] [11], this model utilizes CNNs. The CT scan combines x-ray images taken from different angles to create cross-sectional images. This method with holistic images is significantly different from previous image

patch-based algorithms. Currently, it seems the state of the art method is using CNNs and some form of image preprocessing. CNNs are more effective at ignoring background noise, and image preprocessing increases the visibility of abnormalities. Like [7] and [9], we utilize CNNs, however we do not have a form of image preprocessing to enhance possible areas of abnormalities.

3 Data Set

For this project we used two data sets. The first data set is consisted of 5,863 X-Ray images and 2 categories (Pneumonia/Normal), 4273 (73%) images with pneumonia, and 1583 normal lung images (27%). According to the data set documentation[2], the Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. The dataset was organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). These we divided into 3456 images for training (60%), 1155 images for validation (20%), and 1155 images for testing (20%).

The second data set we used was the NIH Chest X-ray Dataset comprised of 112,120 X-ray images with disease labels from 30,805 unique patients [1]. The data set is divided into 15 classes: There are (14 diseases, and one for "No findings"). Images we classified as "No findings" or one or more disease classes. 60360 (~54%) images were normal while 51760 (~46%) were abnormal. We did not use all the images from this data set; we used two randomly sampled samples. The first batch contained 40,000 images of lung X-rays, 18447 (~46%) images were abnormal and 21553 (~54%) images were normal. These were randomly divided into 36,000 images for training (90%), 2000 images for validation (5%) and 2000 images for testing (5%). The second sample also contained 40,000 images of lung X-rays, 18,317 (~46%) images were abnormal and 21683 (~54%) images were normal. These were randomly divided into 36,000 images for training (90%), 2000 images for validation (5%) and 2000 images for testing (5%).

The third data set we used was the Montgomery County X-ray Set [6]. X-ray images in this data set have been acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. This set contains 138 posterior-anterior x-rays, of which 80 x-rays are normal and 58 x-rays are abnormal with manifestations of tuberculosis. This set we used exclusively for testing.

4 Method

To summarize, most of the effort for this project was spent on developing the model to detect tuberculosis from chest x-ray images. In both of our models we fine-tuned a pre-trained ResNet50 model by freezing all but the last four layers, and adding our own dense layers. Our first approach was to train the model on the small pneumonia data set. The second approach involved training on a much larger data set of abnormal and normal lungs. In addition to using these different CNN architectures, we attempted to optimize detection accuracy by altering the ratio of images in our test/validation/training datasets. Furthermore we tried to, among other things, alter the amount of data and number of epochs to train on. We saved these trained models, then used them to detect tuberculosis.

For both of our models we used the Keras the ResNet50 model.[12] ResNet-50 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 50 layers deep and can classify images into 1000 object categories, and as a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224-by-224. As such we then take advantage of this model and fine-tune it by freezing all but the last four layers, and training our own dense layers. The model resulted in ~24 million trainable parameters.

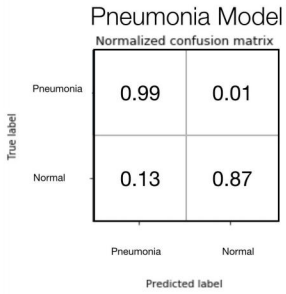
For our loss function we used the cross-entropy loss:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

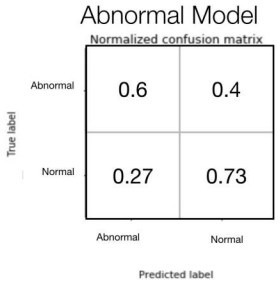
Furthermore we used an Adam optimizer with learning rate of 0.001 and decay 0.00005, and batch size of 10 for all subsets of data. Before training both models we performed some light data augmentation by editing, rescaling by a constant factor, adjusting the shear intensity, doing some zooming, horizontal flipping and some rotation. For the first model we simply used the data set described described in part 3 on 30 epochs before convergence and early stopping after which we obtained good results with respect to test accuracy and then went ahead and tested on tuberculosis. For the second model we initially trained on 15 epochs and obtained a training accuracy, after which we saved the model, and tested it on the tuberculosis set it had never seen before. Call this model Abnormal-A since it trained exclusively on pictures of pictures of normal and abnormal lungs. We then proceed to train a second version of this model, call it Abnormal-B, where we simply trained Abnormal-A for another 15 epochs on the same data and tested on tuberculosis. Lastly Abnormal-C, consisted of continuing training for Abnormal-B, and doing an additional 30 epochs but this time with never before seen data containing new 40000 images.

5 Results and Analysis

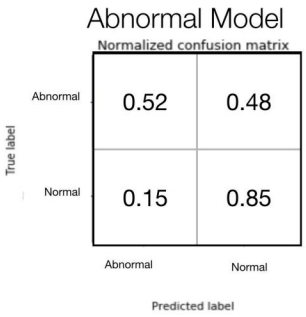
The model which we trained to detect pneumonia performed very well when trying to differentiate between a healthy lung and a lung with pneumonia. The model trained on pneumonia achieved 96% accuracy classifying pneumonia. However, the model did not perform very well with 45% accuracy when classifying tuberculosis. The reason for this can be deduced when looking at the confusion matrix for which it classified pneumonia: Namely what we can observe is that whenever we it was given an image of pneumonia the model very accurately predicted that the lung in the image indeed had pneumonia, it did so with 99% accuracy. However when it came to predicting that a lung was normal it struggled more with only 87% accuracy. We conjectured that the reason for this is relating to the distribution of the data set. As mentioned previously the data on which this model was trained contained almost twice as many pneumonia images as it did normal images, which led us to believe that what our model did was, rather than be able to differentiate between a healthy and a lung with pneumonia, it memorized what a lung with pneumonia looked like. Then when it was asked to differentiate between a healthy lung and one with a non-pneumonia disease, it performed worse than chance.



This is why we decided to train on a larger data set with a better ratio of healthy to unhealthy lungs. Moreover, this is why we decided to train on a data set of healthy lungs in comparison to not simply one disease but rather 14 different diseases (none of which were tuberculosis) all grouped into one: so as to create a more generalizable data set. Abnormal-A only managed ~67.55% testing accuracy, lead negligibly by Abnormal-B at ~67.7% testing accuracy, both of which were beaten by Abnormal-C at a marginally better ~68.75% accuracy. Notable results came about when looking at how each of them performed on the tuberculosis test set. Despite never having seen tuberculosis before Abnormal-A, Abnormal-B and Abnormal-C all outperformed the pneumonia model at 71%, 69%, and 66% accuracy. Because the confusion matrices distribution of accuracies is analogous let us consider Abnormal-A without loss of generality, to understand what happened.



As we can see in the confusion matrix the model was better at correctly predicting that healthy lungs were indeed normal, while it performed a bit worse predicting that an abnormal lung was abnormal. This is likely due to the fact that it was able to build a more solid model of what a healthy lung looked like whereas the "abnormal" lungs were partitioned into 14 different categories. As such we hypothesized that since the network is essentially trained to be able to tell if a lung is healthy or not regardless of what the abnormalities in the lung are caused by, it would be able to better classify an image as having tuberculosis so long as it knows that it was tuberculosis it was testing for. This turned out to be true as seen in the following confusion matrix for when testing on TB.



As we can see the model more accurately identified when a lung was actually healthy than when it actually had tuberculosis, as expected since it was able to develop a model for healthy lungs. For reference here are the model evaluation metrics when testing Abnormal-A on TB:

Classification Report	precision	recall	f1-score	support
NO FINDING	0.71	0.52	0.60	58
ABNORMAL	0.71	0.85	0.77	80
micro avg	0.71	0.71	0.71	138
macro avg	0.71	0.68	0.69	138
weighted avg	0.71	0.71	0.70	138

6 Conclusion and Future Work

Overall, this project was a rewarding deep learning experience, through which we were able to explore methods of developing CNN models, tuning hyperparameters, training and testing our models, and partitioning data sets. In the end, we were able to create a model that was trained on various diseases not including tuberculosis and we were able to detect the presence of tuberculosis on unlabeled images with an accuracy of 71%. This model trained on various diseases outperformed our model trained on pneumonia for detecting tuberculosis. We think this is because the model trained on various diseases was able to generalize better because it learned what normal lungs looked like versus abnormal lungs, where as the model trained on pneumonia was much more constricted and was only reliable when detecting if pneumonia was present.

In the future, as a result of literature review, we learned that image preprocessing is an important step that most disease detection CNNs use. While data augmentation was conducted in order to increase the size of the training set, data preprocessing was not done for the project. For disease detection CNNs, data preprocessing utilizes some sort of method that can enhance sections of the image that have a high probability of an abnormality. Data preprocessing can be conducted on sections of the body where the disease is most likely to appear or manifest. Therefore, using data preprocessing to focus on certain sections may pronounce abnormalities and improve learning for the CNN. Additionally, at least 40 hours was spent on training our model. Finally, at least 60% of the time working on this project was spent learning how to use the GCloud environment. As a result, we did not get to test and tune as many hyperparameters as we would have liked. In the future, we hope to improve our model with further training and hyper parameter tuning to improve accuracies for both testing and tuberculosis detection.

7 Contributions and Acknowledgements

Both Pablo and Eugene worked together on the project the entire time. Developing the model, training and testing, writing the report, and creating the poster was all done as a pair at the same time. We would like to give a special thanks to Aarti Bagul, our project TA, for being a resource to answer questions that we have and giving us insightful feedback. We would also like to thank Kivalu Ramanlal who was our project partner for CS 221 (the project for which this project shares common infrastructure for lung X-ray classification) and who helped develop some of the initial code for this project.

10 Codebase

This code for this project can be found at: https://github.com/etian511

11 References

If a reference is not mentioned in the paper it is because we used it as literature review in the development and study prior to building our network and feel we should cite it.

[1] "NIH Chest X-rays," Kaggle. [Online]. Available: https://www.kaggle.com/nih-chest-xrays/data. [Accessed: 10-Jun-2019].
[2] Kaggle.com. (2019). Chest X-Ray Images (Pneumonia). [online] Available at: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia [Accessed 10 Jun. 2019].
[2] Y. Tian, "Detecting Pneumonia with Deep Learning," Becoming Human: Artificial Intelligence Magazine, 03-Jun-2018. [Online]. Available: https://becominghuman.ai/detecting-pneumonia-with-deep-learning-3cf49b640c14. [Accessed: 02-Jun-2019].
[3] M. Peixeiro, "How to Improve a Neural Network With Regularization," Towards Data Science, 12-Mar-2019. [Online]. Available: https://towardsdatascience.com/how-to-improve-a-neural-network-with-regularization-8a18ecda9fe3. [Accessed: 02-Jun-2019].
[4] R. H. Abiyev and M. K. S. Maaitah, "Deep Convolutional Neural Networks for Chest Diseases Detection," Journal of healthcare engineering, 01-Aug-2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6093039/. [Accessed: 02-Jun-2019].
[5] Bhat, V. (2019). Pneumonia detection ResNet50/VGG16 | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/vnbhat/pneumonia-detection-resnet50-vgg16 [Accessed 2 Jun. 2019].
[6] "Tuberculosis Chest X-ray Image Data Sets - Communications Engineering Branch," U.S. National Library of Medicine. [Online]. Available at: https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets/ [Accessed: 10-Jun-2019]
[7]Lo, S-CB, et al. "Artificial convolution neural network techniques and applications for lung nodule detection." *IEEE Transactions on Medical Imaging* 14.4 (1995): 711-718.
[8]Shih-Chung, B. Lo, et al. "Automatic lung nodule detection using profile matching and back-propagation neural network techniques." *Journal of Digital Imaging* 6.1 (1993): 48-54.
[9]Gao, Mingchen, et al. "Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.1 (2018): 1-6.

[10]Anirudh, Rushil, et al. "Lung nodule detection using 3D convolutional neural networks trained on weakly labeled data." *Medical Imaging 2016: Computer-Aided Diagnosis*. Vol. 9785. International Society for Optics and Photonics, 2016.
[11] Lopes, U. K., and João Francisco Valiati. "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection." *Computers in biology and medicine* 89 (2017): 135-143.
[12] He, K., Zhang, X., Ren, S. and Sun, J. (2019). Deep Residual Learning for Image Recognition. [online] arXiv.org. Available at: https://arxiv.org/abs/1512.03385 [Accessed 10 Jun. 2019].