A Transfer Learning Approach to Pneumonia Diagnosis from Chest X-Rays

Eric Dong, Andrew Castle, Michal Kolakowski

Abstract

Pneumonia is a treatable bacterial, viral, or fungal infection that causes the air sacs in the lungs to fill up with fluid (6). About 1 million people contract pneumonia each year in the United States, with about 50,000 people dying from the disease (1). Accurate and reliable diagnoses of the infection is imperative to preventing negative clinical outcomes, especially because different forms of pneumonia require different types of treatment. For our final project, we attempt to address this issue by developing an image-based deep learning framework to accurately classify pneumonia diagnoses from pre-labeled chest x-ray images (4). We experiment with multiple approaches that leverage transfer learning techniques to adapt pretrained convolutional neural networks (CNNs) towards our dataset (3). We compare our models against naive classifiers and other techniques that do not utilize deep learning. Our deep learning models achieve superb performance when evaluated using accuracy and F_1 score. In the future, we would like to consider multiclass classification to also distinguish between the different pneumonia variants and perhaps even other diseases. Our hope is for our model to improve and expedite the diagnosis process for pneumonia.

1 Introduction

Pneumonia, despite affecting about 1 million US citizens every year, is quite difficult to diagnose. Physicians have several options available to them, but no one method is perfect. Chest x-rays are a common next step after a physical examination reveals signs of the disease. However, doctors are far from perfect at detecting pneumonia from visual inspection of these x-rays. One recent study, for instance, found diagnoses to only be about 67 percent accurate after an x-ray (7). This is compounded by the fact that there are multiple variants of pneumonia (i.e. bacterial, viral, or fungal) which require different treatments. Therefore, not only does the doctor need to accurately determine if the patient has the disease, but he or she must also distinguish between the subtle visual cues in the x-ray to diagnose which kind of pneumonia.

There is clearly substantial room for improvement in the pneumonia diagnosis process. In this paper, we develop a transfer learning framework to classify pneumonia from chest x-ray images. Our system results in far greater accuracy than that of a trained doctor.

2 Dataset and Features

2.1 Dataset

For this project, we obtained gray-scale chest x-ray data from Kaggle (4). The data consists of chest x-ray images depicting the lungs of healthy patients and patients with pneumonia. The specific form of pneumonia is also provided. The data was labeled by two expert physicians and received initial screening for quality control before being posted on Kaggle. The dataset contained a total of 5856 x-rays, pre-separated into 5216 training images, 16 validation images, and 624 testing images.

2.2 Data Processing

The images are quite well organized and standardized: the x-rays are all shot in a similar way and include the same parts of the chest cavity centered in the image. However, the images are not all of the same dimension. Additionally, the raw images

contain a letter "R" indicating the right side of the patient.

To address this, we first re-size each image so that its shortest edge is 256 pixels while maintaining the original aspect ratio. Then we retrieve the center 224x224 pixel portion of the re-sized image. Figure 1 illustrates this process. Each image is then ravelled into a 1-dimensional array of pixels so that each row in our dataset represents an observation with 50,176 features. We use a tensor representation of this dataset for our CNN models and a matrix representation for the other comparative models. We also experiment with training a CNN model on an augmented dataset in which we randomize the re-sizing and cropping process and also perform a random horizontal flip.

Figure 1: Image standardization process

3 Exploratory Data Analysis

3.1 X-Ray Examples

Our dataset is imbalanced towards pneumonia patients over normal patients. Our training, validation, and testing sets contain 3875 pneumonia and 1341 normal cases, 8 pneumonia and 8 normal cases, and 390 pneumonia and 234 normal cases, respectively.

It can be extremely difficult to distinguish an unhealthy x-ray from a healthy one. As Figure 2 shows, to the untrained eye, there is little difference between the two images. In fact, even a highly trained eye will often struggle to identify pneumonia (2) (7). As such, we hope that our image recognition system will be able to extract patterns that the human eye cannot.

3.2 Principal Component Analysis

Principal component analysis (PCA) is a common technique applied to image data. We performed PCA on the entire dataset to derive "eigenlungs" (9) that capture the variation in the dataset. Figure 3 shows the first 25 principal components. Observe that the first few principle components seem to represent the general outline of the chest cavity

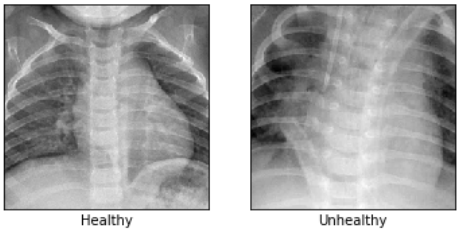

Figure 2: X-ray of healthy lungs (left) and unhealthy lungs (right)

and spine - this feature explains most of the variation in the images. On the other hand, the 21st through 25th principal components seem to represent the rib cage.

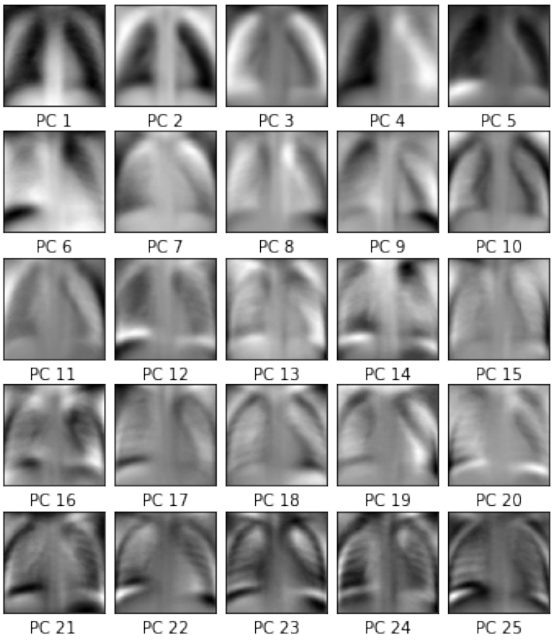

Figure 3: Principal components of chest x-rays

Beyond using PCA for exploratory data analysis, we also leverage this technique as a tool for dimensionality reduction on the data to which we fit our non-deep learning comparative models. From Figure 4, we can see that the first 150 components explain approximately 90% of the variance in the data. As a result, we obtain the first 150 principal components from the training data and transform the entire dataset using these principal components so that each observation now only contains 150 predictors.

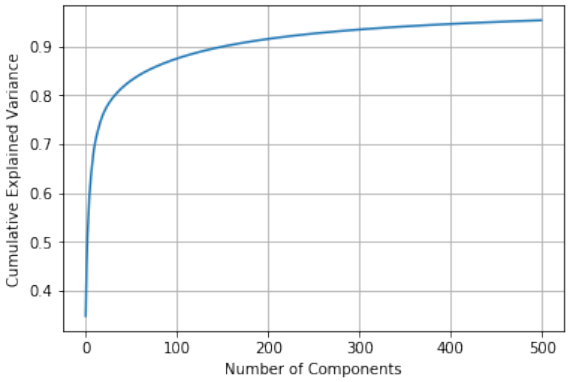

Figure 4: Cumulative explained variance of principal components

4 Methodology

4.1 CNNs and Transfer Learning

Image recognition is a notoriously challenging problem, and one in which substantial progress has already been made. CNNs are well known to be adept at classifying images by filtering, or convolving over the data in order to parse out subtle patterns and spatial relations. However, a CNN capable of recognizing images well often requires millions of parameters and an enormous number of hidden layers, with expensive time and space costs for training. As such, it makes little sense to initialize and train a blank CNN on our relatively small dataset. Such a classifier built on only 5216 images may not be able to learn much of anything.

Instead, we can utilize a technique called transfer learning in which we apply a pre-trained network towards our dataset. By taking an existing network that has been trained on a very large dataset and has already learned patterns important for classification, we can simply re-train the final predictions being made to be on our classification task of interest. Compared to training an entirely new CNN, this last step of re-training is extremely cheap computationally.

The famous ImageNet challenge (8) saw many research teams do the hard work of developing and training state-of-the-art image-based CNN classifiers. These models were trained on over 14 million images with 20,000 categories and achieved exceptionally high classification accuracy. Because the models from ImageNet lend themselves well to transfer learning, we can build on top of that work by specializing a pre-trained network from this challenge to fit our problem.

Specifically, we chose to use the VGG16 model with batch normalization. VGG16 was developed by the Visual Geometry Group at Oxford as a solution to the ImageNet challenge and achieved a top-5 score in the competition.

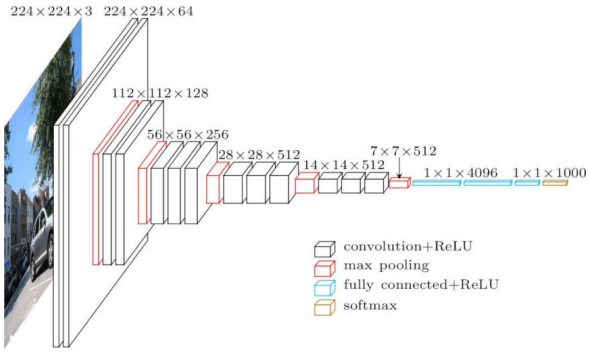

Figure 5: VGG16 architecture

Figure 5 describes the architecture of the VGG16 model. The network contains three fully connected layers following a stack of convolutional layers. VGG16 with batch normalization is identical except it contains a batch normalization layer after each of the VGG16 convolutional layers. For the original ImageNet challenge, images were labeled with one of 1000 different classes: as such, the final layer in VGG16 contained 1000 output channels. For our transfer learning purposes, we froze all of the pre-trained weights of each layer except for this last layer, which we re-trained using our dataset for binary classification of pneumonia.

As mentioned, for our models we used VGG16 with batch normalization with all weights except for those in the final layer initialized and frozen at their pre-trained values. We modified the final layer so that it output two classes. We then trained the final layer of this model on our training set on GPUs using cross entropy loss with stochastic gradient descent optimization. Our CNNs were trained for 10 epochs, with the iteration with the greatest validation dataset accuracy being saved. We trained one CNN using the regular data and one CNN using the randomly augmented data.

4.2 Comparative Models

To evaluate the performance of our deep learning models against a baseline, we built several models for classifying the x-ray images using the PCA transformed training data from our exploratory data analysis. We considered a mixture of linear classifers and ensembled learners. Specifically,

for linear classifiers we applied logistic regression, hard-margin support vector machines (SVM), and linear discriminant analysis (LDA). Our ensembled learners consisted of random forest (RF) and gradient boosting (GB). Our RF classifier used 500 trees with a maximum depth of 5 nodes, split on 12 features based on information gain. And our GB classifier optimized deviance with a learning rate of 0.1 with 100 boosting stages. Note that while we trained each of these models on the training set, we did not use the validation data nor did we cross validate to tune the parameters.

5 Results

Overall, each of our models achieve a higher testing accuracy than that of trained doctors as well as naive classifiers (i.e. predicting all pneumonia). Table 1 summarizes the performance of each model we considered.

Model	Test Accuracy	F_1 Score
CNN w/ regular data	0.8782	0.9106
CNN w/ augmented data	0.9455	0.9578
Logistic Regression	0.7596	0.8366
SVM	0.7660	0.8371
LDA	0.7756	0.8458
RF	0.6474	0.7796
GB	0.7660	0.8385
All Healthy	0.3750	–
All Pneumonia	0.6250	0.7692

Table 1: Model performance

Our final model, the CNN using augmented data, is extremely accurate in predicting pneumonia from chest x-rays. When used on test data, the model shows a very impressive 94.55% accuracy. Additionally we managed a recall of 99.0% and a specificity of 92.8% for a total F_1 score of 0.9578, as illustrated in Table 2. The current top F_1 score on Kaggle seems to be 0.875, so our model represents quite an improvement from the existing standard (5).

Random forest, gradient boosting, logistic regression, and support vector classification all performed reasonably well. The best comparative model we created, however, was our linear discriminant analysis model which acheived a 77.56% test accuracy score. This is comfortably higher than the $\sim 67\%$ accuracy of a trained radiologist.

	Predicted Healthy	Predicted Pneumonia
Actual Healthy	204	30
Actual Pneumonia	4	386

Table 2: Confusion matrix of CNN with augmented data

6 Conclusion

Our transfer learning application results in an extremely accurate model for pneumonia detection. From a chest x-ray alone, our model is able to diagnose pneumonia far more accurately than experienced radiologists. The final deep-learning model also significantly outperforms more parsimonious models that build upon principal component analysis, justifying the use of a more complicated, black-box classifier. Furthermore, our CNN model has remarkably high recall, which is crucial in an application like pneumonia detection where the costs of misclassifying a sick patient can be deadly.

We are able to obtain these impressive results because of the way that we have leveraged past ImageNet work: by taking advantage of models from Imagenet that have trained for weeks, we can build a highly specialized application off of a relatively small amount of data. The VGG16 architecture decomposes our images down into key features that we can then use as predictors of lung health.

While there are other sophisticated lung health models on the market, such as CheXNet (10), such systems require a massive amount of x-ray data to function. By using transfer learning, we can create a highly effective model with complexity while drawing on only a few thousand observations. This is exciting because it opens the door for many problems and domains with datasets previously thought to be too limited for more complex and powerful models.

7 Future Work

While we are pleased with our results, there are additional extensions and improvements that may be pursued. First, it would be of interest to explore using our transfer learning framework with different CNNs, such as AlexNet, ResNet, and Inception. The ImageNet contest is held annually and each year, more accurate models are developed. Applying these models to our framework would likely yield even better results.

Additionally, because our dataset was relatively limited, it would be interesting to further explore the idea of data augmentation. Data augmentation is generally good practice with image classification and by using a wider variety of transformations, we could perhaps not only improve our models but we could also reserve more of the dataset for testing purposes.

Lastly, it would be useful to build a multiclass classification model that distinguishes the different variants of pneumonia in addition to detecting the disease, or better yet a model that detects multiple diseases. As mentioned before, each form of pneumonia requires a different type of treatment (i.e. antibiotics versus supportive care). A model that is capable of identifying many diseases at once would provide the most value to doctors.

References

[1] Pneumonia Can Be Prevented-Vaccines Can Help — CDC. Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, www.cdc.gov/pneumonia/prevention.html.

[2] Kelsberg, Gary, and Sarah Safranek. "How accurate is the clinical diagnosis of pneumonia?." Clinical Inquiries, 2003 (MU) (2003).

[3] Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." Cell 172.5 (2018): 1122-1131.

[4] Mooney, Paul. Chest X-Ray Images (Pneumonia). Kaggle, 24 Mar. 2018, www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[5] NAIN. Beating Everything with Depthwise Convolution. Kaggle, www.kaggle.com/aakashnain/beating-everything-with-depthwise-convolution.

[6] Pneumonia. National Heart Lung and Blood Institute, U.S. Department of Health and Human Services, www.nhlbi.nih.gov/health-topics/pneumonia.

[7] Nazerian P et al. Accuracy of lung ultrasound for the diagnosis of consolidations when compared to chest computed tomography. Am J Emerg Med 2015 May; 33:620. (http://dx.doi.org/10.1016/j.ajem.2015.01.035)

[8] Olga Russakovsky and Jia Deng and Hao Su and Jonathan Krause and Sanjeev Satheesh and Sean Ma and Zhiheng Huang and Andrej Karpathy and Aditya Khosla and Michael Bernstein and Alexander C. Berg and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision (IJCV). (2015).

[9] Sirovich L, Kirby M. "Low-dimensional procedure for the characterization of human faces." Journal of the Optical Society of America. (1987).

[10] Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv. (2017).