Capstone Design Proposal

8/27/18

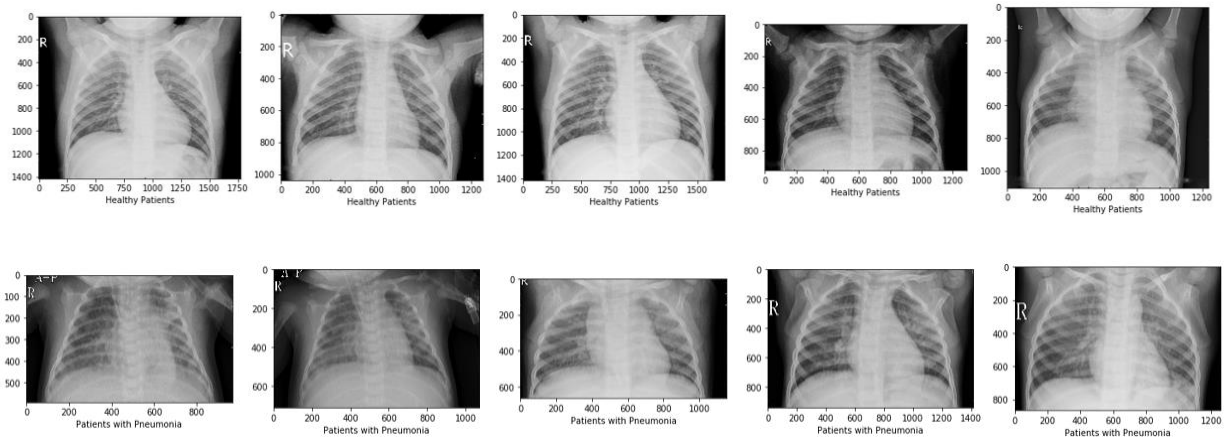
Domain Background

"The Flu is killing up to 4,000 Americans a Week" The Flu also known as influenza is a contagious viral infection that can be spread through coughing or sneezing. Pneumonia is a condition characterized by the inflammation of the lungs. The flu is a common cause of pneumonia, and in 2015 they were the eighth leading cause of death within the United States.

There were 40,414 deaths in the U.S. during the third week of 2018, the most recent data available, and 4,064 were from pneumonia or influenza, according to the CDC data. The number for that week is expected to rise more reports are sent to the agency. [i] Although, additional tests are often used to diagnose this condition X-rays are still one of the most widely used, cost effective methods. Images are taken of patients are then visual analyzed to make a diagnosis or to determine if further evaluation is required.

Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally and in 2015, 920,000 children under the age of 5 died from the disease. While common, accurately diagnosing pneumonia is a tall order. It requires review of a chest radiograph (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. This competition seeks to improve the efficiency and reach of diagnostic services.

It is fascinating to me that when comparing the X-rays of healthy patients with those who have pneumonia that it difficult even notice differences between the examples.



I chose this application because I believe that the successful implementation would save lives, and as a novice machine learning student it would bring me great satisfaction to know that something I worked on provided real value, it gives me confidence that I am able to implement a solution that would have more than just academic merit.

Problem Statement

The goal is to use the labeled image dataset provided to train a Neural Network and the test the accuracy with which the model accurately diagnosis the condition.

Datasets and Inputs[ii]

The dataset titled, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification" is provided by Daniel Kermany via Kaggle. It is organized into 3 folders (train, test, validation) and contains subfolders for each image category (Pneumonia/Normal). There are 5,860 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). Of that total 1,585 files are of patients in good health (Normal), and 4,275 are of patients with various forms of pneumonia all labelled, Pneumonia. The images are different sizes but will be resized into uniform dimensions in order to be processed.

Source: https://data.mendeley.com/datasets/rscbjbr9sj/2

Solution Statement

The goal of this project is to implement transfer learning to train a Convolutional Neural Network using by selecting the architecture that provides the best accuracy based on the given test set.

Benchmark Model

The following results were recorded describing the effectiveness of pneumonia using X-ray:

The initial chest radiographs of 31 patients with laboratory-proved pneumonia were evaluated by a panel of 6 radiologists who had no prior knowledge of the clinical data. No statistical reliability was found for distinguishing bacterial from nonbacterial pneumonia. Radiographic diagnoses were 67% accurate for the 16 cases of bacterial pneumonia, and 65% accurate for the 9 viral cases.[iii]

An accuracy of 60% or better would be result in an improvement compared to the study referenced.

Evaluation Metrics

The value of the model will be evaluated using the accuracy metric. The accuracy is defined as the number of times the model correctly predicts the same value as the actual value in the data set divided by the number of attempts. In this case, accuracy will be used to make a comparison to the benchmark results set forth in the paper above. However, this use case also suggest that in practice the measure recall may be an even more effective metric. Recall is defined as the number of cases that the model correctly predicts, true positives divided by the sum of true positives and false negatives. False negatives in this example would be people with pneumonia who are diagnosed as healthy. Conversely, true negatives would be health patients who were told that they have pneumonia. While the latter case may be scary proposition, and less than ideal, the former case is deadly. A sick patient misdiagnosed as healthy is a more important classification error and therefore a model that maximizes recall should probably be used in practice. The recall score will be calculated to evaluate the performance of the model.

Project Design

The project can be broken down into these discrete steps:

- Load data and visualize images
- Pre-processing/ Rescaling Images
- Break dataset into Training, Testing and Validation sets
- Define the Model Architecture
- Compile the Model
- Train the Model
- Load the model weights with best validation accuracy
- Calculate the classification accuracy on the Test Set

The most significant portion of time should will be dedicated to defining the model architecture. This is an iterative step that will require decisions to be made about the number of layer, types of layers and other associated hyper-parameters needed to achieve the best possible prediction accuracy on the test set.

---

[i] http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/pneumonia/what-is-the-connection.html
[ii] https://data.mendeley.com/datasets/rscbjbr9sj/2
[iii] https://pubs.rsna.org/doi/10.1148/124.3.607?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed