Machine Learning Engineer Nanodegree

Capstone Proposal

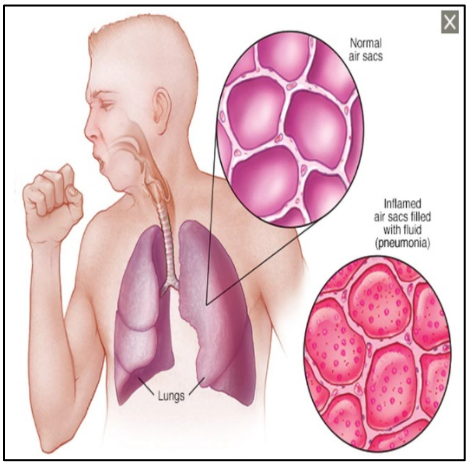
Nitesh Vachhani
September 15th, 2019

Proposal

Domain Background

Pneumonia is a form of acute respiratory infection that affects the lungs. The lungs are made up of small sacs called alveoli, which fill with air when a healthy person breathes. When an individual has pneumonia, the alveoli are filled with pus and fluid, which makes breathing painful and limits oxygen intake.

Pneumonia is the world's leading cause of death among children under 5 years of age, accounting for 16% of all deaths of children under 5 years old killing approximately 2,400 children a day in 2015. There are 120 million episodes of pneumonia per year in children under 5, over 10% of which (14 million) progress to severe episodes. In the US, pneumonia is less often fatal for children, but it is still a big problem. Pneumonia is the #1 most common reason for US children to be hospitalized.



There were an estimated 880,000 deaths from pneumonia in children under the age of five in 2016. Most were less than 2 years of age.

Diagnosing pneumonia early is key to faster and healthy recovery. Pneumonia is typically diagnosed based on a combination of physical signs and a chest X-ray. Pneumonia usually manifests as an area or areas of increased opacity in an X-ray. However, the diagnosis of pneumonia from x-ray is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, etc.

In developing countries like India , rural health services is a serious problem, with a shortage of qualified healthcare providers as a major cause of the unavailability and low quality of healthcare. Applying image-based machine diagnostic medical techniques could improve healthcare outcomes in rural areas of developing countries.

Citation of work in this field: https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5

.

Problem Statement

Applying image processing techniques on medical images can offer an objective opinion to improve efficiency, reliability, and accuracy in the medical diagnostics process.

The problem being targeted for the capstone project is to **build a machine learning model to detect if a child is suffering from pneumonia using chest x-ray images**.

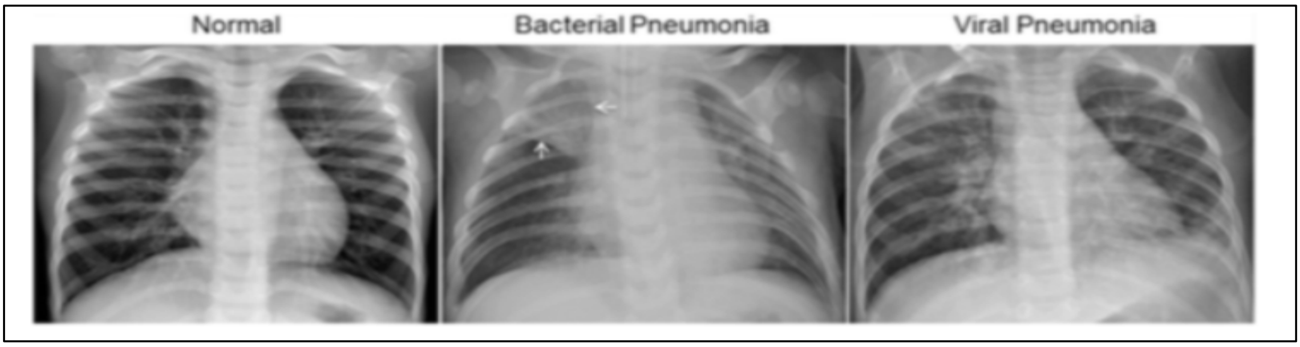
Datasets and Inputs

The dataset is basically a set of chest x-ray images. The data set is taken from Kaggle and is available here - https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

Below are few of the important details about this dataset.

- The dataset contains 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). The dataset itself is organized into 3 folders of train, test and validate.
- Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of paediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care.
- For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being added to the dataset. In order to account for any grading errors, the evaluation set was also checked by a third expert.

Sample Images:



Solution Statement

To check if the chest x-ray image implies normal or pneumonia, a classification model will be built. Instead of building a simple logistic regression model, the solution will build a Convoluted Neural Network model. Since it's an image classification problem transfer learning technique will be incorporated to the solution to built a more accurate and robust solution. The model will then be validated against the test set and the results will be evaluated.

Benchmark Model

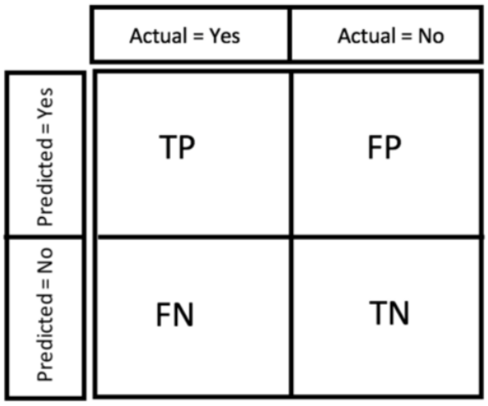
For this problem of image classification, logistic regression model will be created and that will be considered as benchmark model.

The logistic regression model will be compared against a CNN model and solution model of transfer learning with CNN. The benchmark model will be compared with the solution on precision, recall and confusion matrix (details in next section) to check how well the model performs as compared to the benchmark.

Evaluation Metrics

Both the benchmark model and the solution model will be evaluated using the confusion matrix. More specifically, precision and recall for the models will be calculated which will serve as the evaluation metrics.

Confusion Matrix - A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.



	Actual = Yes	Actual = No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

- **True Positives (TP):** when the actual class of the data point was 1(True) and the predicted is also 1(True)

- **True Negatives (TN):** when the actual class of the data point was 0(False) and the predicted is also 0(False)

- **False Positives (FP):** when the actual class of the data point was 0(False) and the predicted is 1(True).

- **False Negatives (FN):** When the actual class of the data point was 1(True) and the predicted is 0(False).

Recall - Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).
Recall is given by the relation:

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Precision- To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).
Precision is given by the relation:

$$\textit{Precision} = \frac{TP}{TP + FP}$$

Project Design

Below are the workflow steps which will be carried out as part of building a solution

1. **Loading & Visualizing the Data** – The images from different folders will be loaded and images will be viewed to get a better understanding of the data at hand. The number of images and the ratio of normal/ pneumonia in each of the folders will be checked to understand the distribution.

2. **Data Pre-processing/Augmentation** – Data augmentation techniques will be applied as a pre-processing step which will enable us to significantly increase the diversity of data available for training models, without actually collecting new data.

3. **Baseline Model Processing**
   a. Prepare data for a standard CNN Model.
   b. Train the model.
   c. Validate the model against the test set
   d. Generate metrices using evaluation criteria.

4. **Solution Model Processing**
   a. Prepare the data for a CNN Transfer Learning Model
   b. Select the base model (Eg. VGG16 , Inception , Resnet50 etc.)
   c. Apply own fully connected layers and activation functions
   d. Compile and Train the model
   e. Validate the model against the test set
   f. Tune parameters and try steps b to e till satisfactory model is obtained.
   g. Generate metrics using evaluation criteria.

5. **Compare Metrices –** Compare the evaluation metrices for the baseline model and the solution model.

6. **Conclusion** – Report conclusion for the capstone project.