Udacity Machine Learning Engineer Nanodegree

Capstone Proposal: Differentiate Between Chest X-Rays with Pneumonia and Chest X-Rays Without Pneumonia Using a Convolutional Neural Network (CNN)

Oliver Dixon
September 21, 2020

Domain Background

Image Classification is being developed and used extensively in the medical industry in order to more accurately diagnose illnesses. X-rays and MRI scans are being used to identify Pneumonia in the case of this project to identifying the onset of dementia and Alzheimer's in brain MRI scans. The applications of machine learning and AI in the medical field are vast and could eventually lead to a type of Star Trek scanner that can diagnose someone by means of image classification. The popularity of this field can be seen in the amount of research and development that has gone into Covid-19 using machine learning methods.

Problem Statement

There are a few problems generally that this technology can solve. Doctors can make incorrect diagnoses or there might be a lack of resources. For example, if a good enough CNN model existed to diagnose people with Covid-19 from X-rays there would not be the problem of testing we see today. In some developing countries, resources are limited and there might not be doctors to serve some sectors of the community. However, if there was a machine working 24 hours a day then people could walk through an X-ray machine and immediately get a diagnosis and get a prescription or medicine almost immediately. Even in developed countries, sometimes people have to wait weeks even before they can see a doctor. The development of this technology would hopefully solve problems in the developing world and the developed world.

Datasets and Inputs

The dataset is available on Kaggle. Chest X-ray images of children ranging from one to five years were made available from Guangzhou Women and Children's Medical Center in Guangzhou, China. The dataset is organized into three folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-ray images. Low quality or unreasonable scans were removed. The diagnoses for the images were graded by two expert physicians before being cleared for training for an AI system.

Solution Statement

A CNN will be developed and trained using Tensorflow in AWS Sagemaker. Image augmentation will be utilized in order to improve results. Sagemaker's Automatic Model Tuning will be used for the best results in hyperparameter tuning.

Benchmark Model

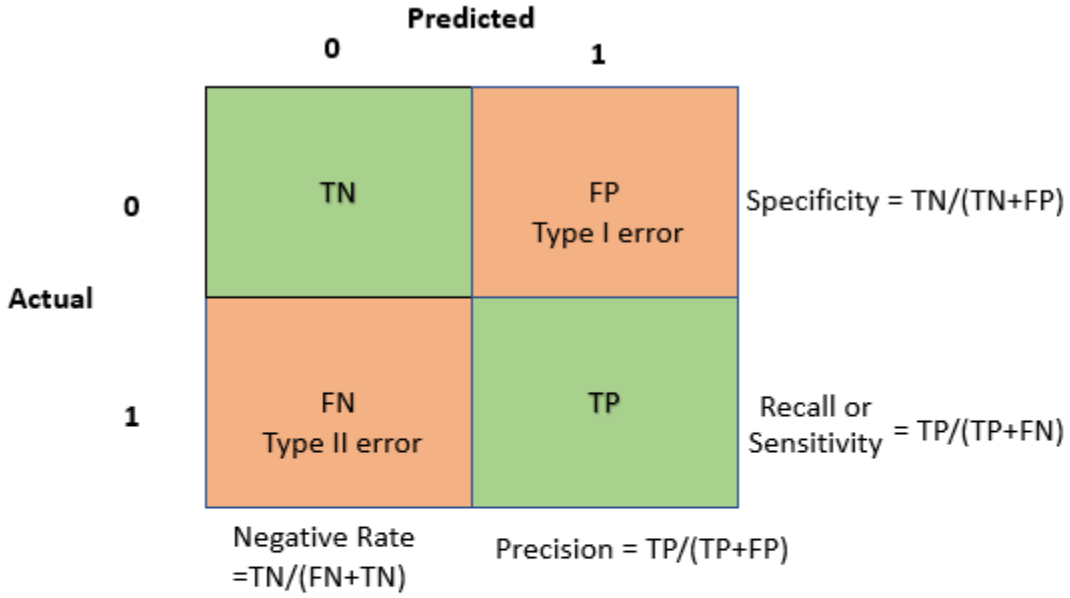
One of the highest ranked notebooks on Kaggle using Tensorflow that seemed the most complete and had the most positive comments had the following results:

Recall: 0.9799	Val_Recall: 0.9788	Test_Recall: 0.9974
Precision: 0.9966	Val_Precision: 0.9906	Test_Precision: 0.7466
Accuracy: 0.9775	Val_Accuracy: 09775	Test_Accuracy: 0.7869

The recall and accuracy will be used to measure how well my model is performing in light of what seems to be possible.

Evaluation Metrics

The CNN generated will be evaluated by creating a confusion matrix in order to see the true Positive(TP) and true negative(TN) against the false positives(FP) and false negatives(FN). Thereafter, the precision, recall and F1 score will be calculated. The recall score is an important score when it comes to diagnosing because the cost of classifying someone not having an illness when they actually have it is much higher than classifying that someone is ill when they are not.



$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{F1 - Score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Project Design

1. Load the Data
2. Visualize the Data
3. Build the CNN

4. Train the Model

5. Deploy the Model
6. Make Predictions and Evaluate

7. Delete Endpoint

8. Hyperparameter Tuning

9. Deploy Best Model

10. Make Predictions and Evaluate

11. Delete endpoint

References

https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

https://www.kaggle.com/amyjang/tensorflow-pneumonia-classification-on-x-rays

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