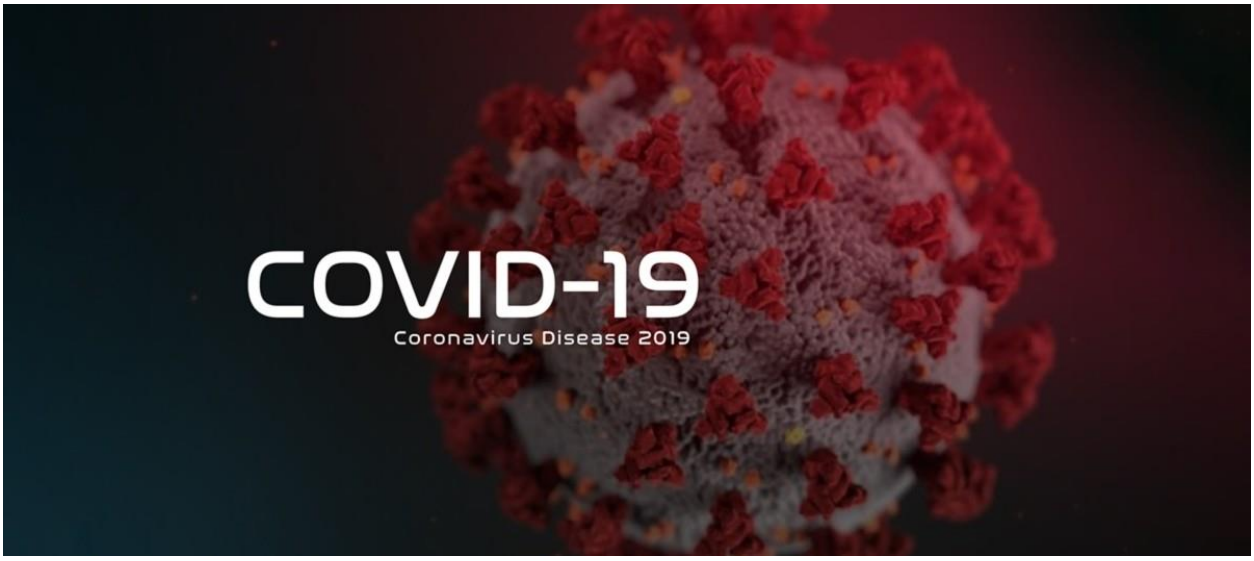
Machine Learning Engineer Nanodegree

Capstone Proposal

Detecting Covid-19 in x-ray Images

Mohammed Lokma

6/8/2020



Domain Background

These days, the world has woken up to the Covid-19 pandemic

Which is known as corona virus

The COVID-19 coronavirus is the global health crisis of our time and the biggest challenge we have faced since World War II. Since it appeared in china late last year, the virus has spread to every country except Antarctica

Not only the a health crisis, just as it is an unprecedented social and economic crisis. By emphasizing each country that touches it, this creates a devastating social state, The economic and political implications that will leave deep and long-term scars. The United Nations Development Program is the technical lead in the United Nations' social and economic recovery, along with the health response led by the World Health Organization and the Global Humanitarian Response Plan, and working under the leadership of the United Nations resident coordinators.

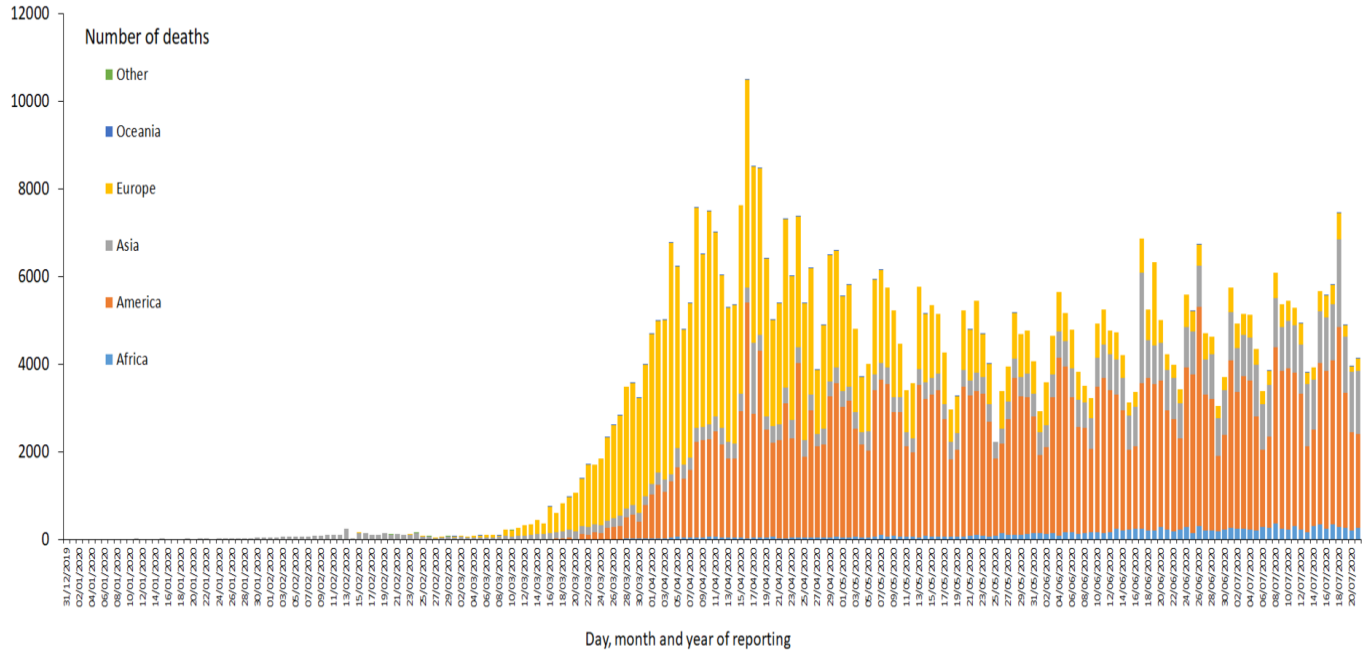
Statics

Let us see the covid-19’s effect of economic

Every day, people lose their jobs and income, and there is no way to know when normal life will return. Small island states, which rely heavily on tourism, have empty hotels and deserted beaches. The International Labor Organization estimates that 195 million jobs can be lost.

The World Bank predicts a drop in remittances of $ 110 billion this year, which could mean that 800 million people will be unable to meet their basic needs.

Globally, as of **3:37pm CEST, 20 July 2020**, there have been **14,538,094 confirmed cases** of COVID-19, including **607,358 deaths,** reported to WHO.



Problem Statement

Now , corona virus threatens everything . economic, people jobs , people life and our normal life .

Who And most scientists are now trying to find a vaccine or any way to reduce the spread of the pandemic

The disease is currently diagnosed by a hospital swab

According to the World Health Organization and by tracking the pathology within the human body, they discovered that the virus first attacks the respiratory system, so we can know that the person is pregnant or not through the X-ray image of the respiratory system, thus

reducing the cost of detecting the disease and reducing the pressure on the detectors

Solution

The problem here is to predict if a person is has the virus or not, depending on the x-ray scan of the chest.

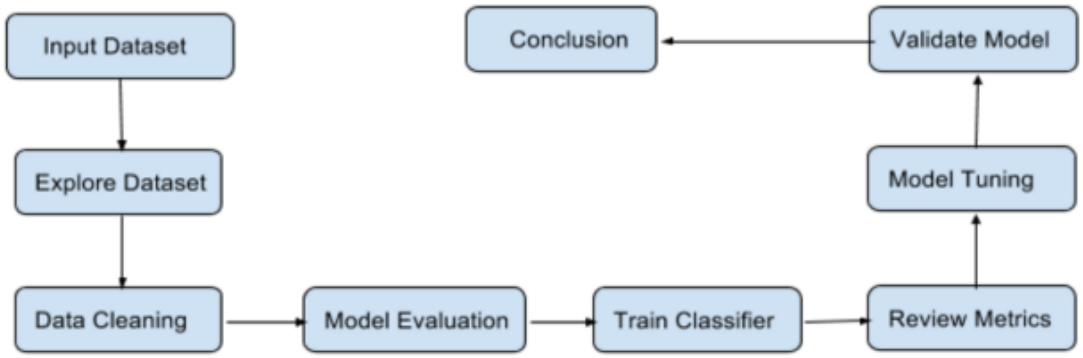
Datasets and Inputs

- I used Dr. Cohen repository (**GitHub repo.**)
- Inside the repo you'll find example of COVID-19 cases, as well as MERS, SARS, and ARDS.

- Covid X-Ray Image Dataset for positive cases (https://github.com/ieee8023/covid-chestxray-dataset)

- Kaggle X-Ray Chest Images for negative cases (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia)

Implementation



Benchmark Model

This is a binary classification model

I will use the Convolutional Neural Networks (CNN) model .

I will create it using transfer learning which has more accuracy rather than the model which is being created from scratch

Evaluation metrics

For this multi class classification ,Multi class log will be used to evaluate the model .

Confusion metrics (What is this ?)

How is the classification being evaluated ?

A confusion matrix is a table often used to describe the performance of a classification model .

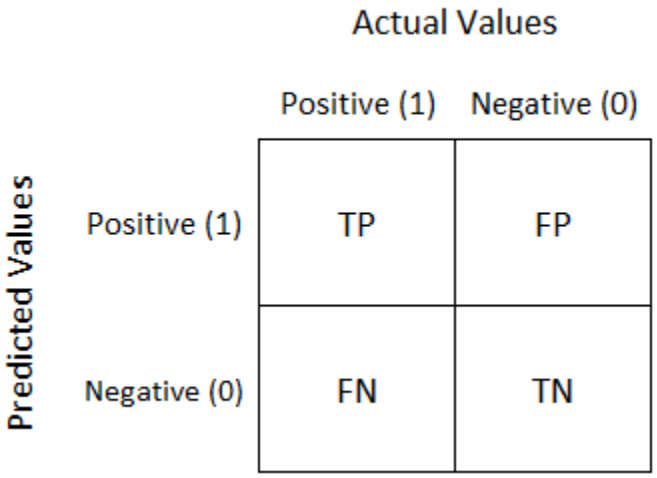
|It is used on a set of test data for which the true values are known.

It allows the visualization of the performance of an algorithm.

If I want to summarize Confusion matrix. It is a summary of prediction results on the classification problem. The number of correct and incorrect predictions are summarized with the count values and are divided according to each category.

The Confusion Matrix shows the ways in which your classification model is confused when it makes predictions. It gives us an insight into not only the mistakes a classifier makes but more importantly the types of errors that are made.

It is a table with 4 different combinations of predicted and actual values.



It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve.

So what is the meaning of TP , FP , FN , TN ?

True Positive:

Interpretation: You predicted positive and it's true.

True Negative:

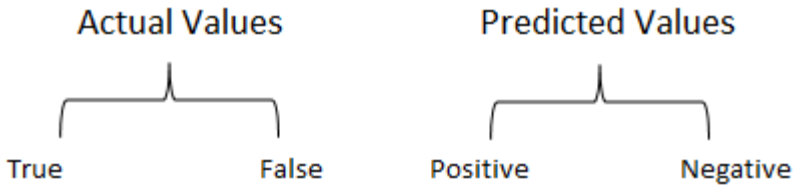
Interpretation: You predicted negative and it's true.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

False Negative: (Type 2 Error)

Interpretation: You predicted negative and it's false.



Classification Accuracy:

Classification Rate or Accuracy is given by the relation:

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

Recall :

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

Recall is the ratio of the total number of correctly classified positive examples divide to the total number of positive examples

Precision

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

Out of all the positive classes we have predicted correctly, how many are actually positive.

High recall, low precision: This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.
Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure:

$$F\text{-}measure = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score

helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

The sensitivity indicates the correctly rated positive rate and is equal to the true positives divided by the sum of TP and FN. sensitivity may be referred to as a true positive rate.

$$\text{Sensitivity} = \frac{TP}{FN + TP}$$

Specificity refers to the rate of correctly classified negative and is equal to the ratio of TN to the sum of TN and FP

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Code : Python code to explain the above explanation

```
# Python script for confusion matrix creation.

from sklearn.metrics import confusion_matrix

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report


actual = [1, 1, 0, 1, 0, 0, 1, 0, 0, 0]

predicted = [1, 0, 0, 1, 0, 0, 1, 1, 1, 0]

results = confusion_matrix(actual, predicted)
```

```
print 'Confusion Matrix :'

print(results)

print 'Accuracy Score :',accuracy_score(actual, predicted)

print 'Report : '

print classification_report(actual, predicted)
```

Project Design

Step 1 : import the necessary libraries

Step 2 : pre-process the data , validate it and create the train , validation and test dataset

Step 3 : Build CNN model

Step 4 : train the model

Step 5 : Deploy the model

Step 6 : test it

References

- https://www.undp.org/content/undp/en/home/coronavirus.html

- https://covid19.who.int/?gclid=Cj0KCQjwpNr4BRDYARIsAADIx9zwAu9ZUMiqtVJ_liv_5VOw3du1f0K_Vg9jjUycN8DASK3SoHvLPScaAqHFEALw_wcB

- https://www.xlstat.com/en/covid

- https://medium.com/@madhavvyas_10553/bank-marketing-campaign-prediction-project-99fe1c3029a4

- https://www.geeksforgeeks.org/confusion-matrix-machine-learning/#:~:text=A%20confusion%20matrix%20is%20a,the%20performance%20of%20an%20algorithm.

- https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62