Identifying COVID-19 from X-ray Images using Convolutional Neural Networks

Ali Arshad

August 8, 2020

Table of Contents

Motivation and Background

Entire cities are under quarantine, the world's economy is suffering and thousands of people are dying. In the midst of this global crisis, nurses, doctors, volunteers and other essential workers are risking their lives on a daily basis. The world is facing its biggest pandemic in decades, and all because of the novel coronavirus disease, also known as COVID-19. While all this is happening the scientific community is coming together to mitigate and fight against the virus. The goal of this project is to spread a little more awareness and knowledge about the virus, as well as, demonstrating the powerful tools that are available in the world of data science.

Problem Statement

This project aims to understand if it is possible to classify the presence of COVID-19 in a patient, solely by looking at one of their chest X-rays. There is a need to test millions of people everyday, which could cost billions of dollars[1][2]. Exploring the idea of using a convolutional neural network as a viable means to identify the disease, may cut these costs while still providing high accuracy diagnostics. If such methods of testing appear to be feasible then there is a chance with further development and research that analysis of medical images can pave the way to a future without this virus.

Datasets

Collection

The project uses two data sets. The first being a collection of 357 X-ray images focusing on patients with COVID-19[3]. Included in the set are images of patients that suffer from diseases that share symptoms with COVID-19, such as MERS, SARS, and ARDS. Each image in the set has its own metadata which adds context to the scan. The metadata contains fields such as, filename, age, sex, view from the image, and the findings. The view describes the angle the image is taken from the person. The findings describe what the patient ended up being diagnosed with. The findings and views will be considered later in the data cleaning process.

The second dataset contains over 10'000 X-ray images[4]. The dataset is partitioned into two parts, scans of patients who are diagnosed with some form of pneumonia and scans of patients who are determined to be healthy. All images from this dataset, unlikes from the first dataset, are framed on the patient's anterior-posterior.

Data Cleaning

As a part of the data cleaning process, the data is partitioned into two sections. The first section is a list of X-ray images of patients diagnosed with COVID-19. The data from this section is pulled directly from the first dataset. The metadata from the first dataset allowed for the images to be separated across each section by their diagnosis. That separation is how we exclusively extracted the COVID-19 related images from the dataset.

The second section is a list of X-ray images of patients who were NOT diagnosed with COVID-19, this includes patients diagnosed with other diseases that affect the lungs, such as SARS and ARDS. The data from this section contains all of the images that were not used from the first dataset. Since there are significantly more images in the first section than the second section, an arbitrarily chosen set of images from the second dataset were added to the second section. The second section was balanced to have a similar amount of images of patients who were healthy versus images of patients who were not healthy.

Our application will only consider images that are X-rays taken from the view of the anterior-posterior or posterior-anterior of the patient. Both sections of the data are then cleaned of images of CT scans (see figure 1). They are both also cleaned of images that are taken from the left of the patient (see figure 2) or from an axial view (see figure 3).

Figure 1: CT scan of a patient with COVID-19

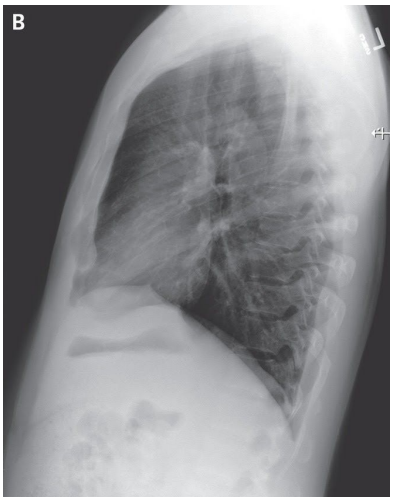

Figure 2: X-ray image viewed from the left of a patient

Figure 3: An axial view of a patient's lungs

The final cleaned section dataset that is compiled contains a total of 505 X-ray images, with 250 images of patients in the COVID-19 section and 255 images in the non-COVID-19 section. Figure 4 is an example of an image that can be expected to be seen in the COVID-19 section of our final dataset. Figure 5 and 6 are an example of images that can be expected to be seen in the non-COVID-19 section of our final dataset.

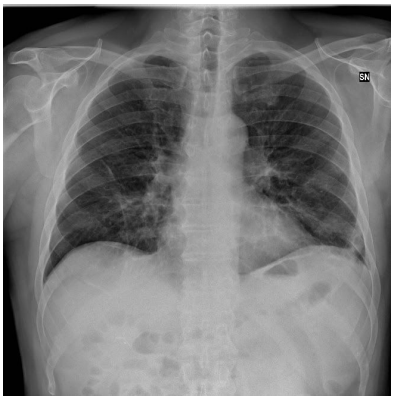

Figure 4: X-ray image of a patient with COVID-19

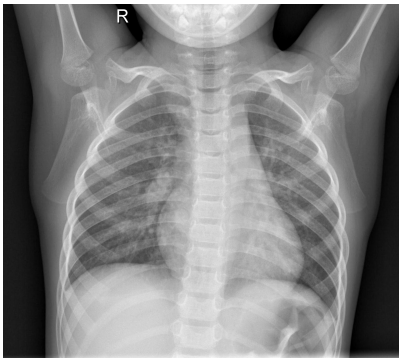

Figure 5: X-ray image of a healthy individual

Figure 6: X-ray image of a patient with a severe case of ARDS

Architecture/Pipeline

The high level project architecture is relatively simple and linear.

1. Retrieve the dataset we compiled to be used as training/testing data.
    a. Label all the images as either ‘covid’ or ‘non-covid’
    b. Resize all the images to the same dimension
    c. Divide the data into a training set and a testing set
2. Build and compile the convolutional neural network model
    a. The neural network starts by using the X-ray image data as an input layer
    b. The network consists of 5 main convolutional block with the image
    c. After the blocks is a FC layer which leads to the output layer
    d. The output layer has 2 variables to indicate, ‘covid’ and ‘non-covid’
3. Train the model on the training data
4. Test the trained model on the testing data
5. Output the results of the testing and save the model

Methodology

This project will be using the data mining techniques of pattern mining, and binary data classification. Pattern mining is utilized by the neural network to understand what sort of patterns are indicative of having COVID-19. Once our model is trained after the pattern mining stage, it will be able to perform binary classification on X-ray images to either label them as having or not having COVID-19. The project structure is based off of another COVID-19 classifier made by Adrian Rosebrock[5].

Images files hold a lot of data, to the human eye, images are discernible by the features that make their whole. Taking a step back, a feature is just a collection of patterns that are arranged in a specific way. It may seem like humans are better at classifying images, but when the differences in features between two classes of images becomes small enough, it can become virtually impossible to rely on human sight.

Looking between X-rays of the patients with COVID-19 and comparing them to non COVID-19 X-rays, it becomes apparent that the differences can be too subtle for humans to reliably notice. This is where convolutional neural networks (NNs) take their throne. NNs can break down classes of images into the key features that build them up. Each of those features is composed of a set of patterns and each of those patterns are composed of even smaller patterns, and so on. Image analysis and pattern mining becomes a simple task for these NNs when it's broken down into such atomic parts.

The NN is built using Python 3 and the Tensorflow library. Tensorflow allows easy access to NN related training tools and compilation. For this project Tensorflow is good because it is general purpose and has ease of use. However, for the most optimal results a NN built from scratch, specifically designed to classify COVID-19 X-rays would work better. The amount of time and development hours needed to build such a NN is outside of the scope of this project. In addition to Tensorflow, we also use the sci-kit and matplotlib libraries to collect and visualize our results.

The layers chosen for this particular NN are based off of a similar neural network by Abhinav Sagar which was used to classify pneumonia in patients[6]. This particular network was chosen for its effectiveness in recognizing patterns from X-ray images. One of the major differences between our networks is our output layer. Our output layer has two nodes instead of one. This allows us to treat that final layer as a binary classifier. One node represents the X-ray showing signs of COVID-19 the other showing the opposite.

Evaluation and Results

Through trial and error in an attempt to maximize accuracy and minimize loss, minor adjustments were made to parameters like the batch size, epochs and the FC layer of the NN. The goal in adjusting these parameters was to avoid overfitting the model to the training data, while still being able to provide a sufficient amount of training. It is also important to consider not to allow your model to have too many variables. In those cases the NN will have enough memory to store specific noise related to the training data; this ends up leading the network to be overfit.

After the adjustments have been made, we have the network that is shown in figure 7.

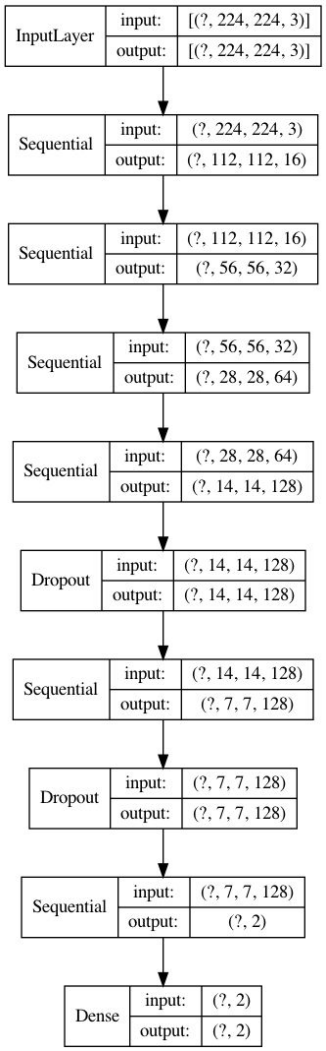

Figure 7: A visualization of the layers that are compiled together to make the convolutional neural network

After analyzing the changes in loss and accuracy while training (figure 8 and 9), we can see the rate of training plateauing. This indicates that with additional epochs, the accuracy of the model will change by a negligible amount.

Figure 8

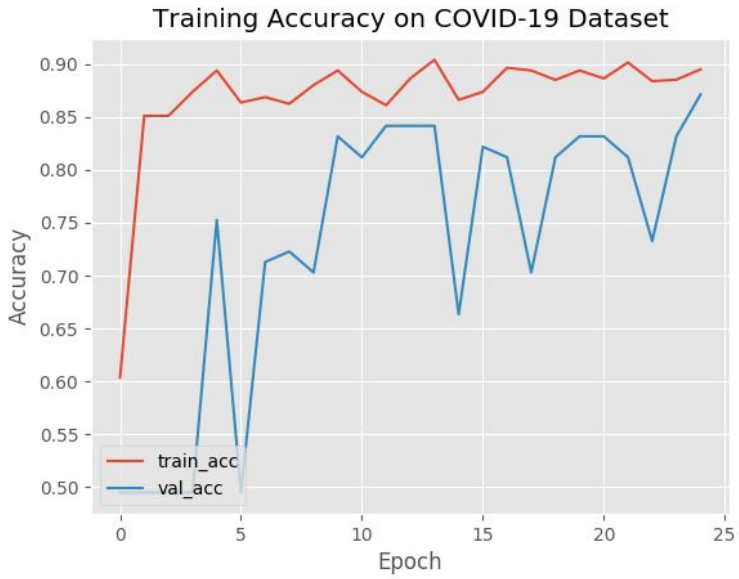

Figure 9

Once we evaluate our model on our testing data we are left with the results in tables 1, the relatively consistent values in precision and recall show us that we have a balance between false positives and false negatives. Additionally the data from table 2 supports these findings. An accuracy of 87.13% is a good start when identifying the disease. These X-rays look nearly identical yet our network was able to differentiate between them. However, on the scale of hundreds of millions of people, this model would lead to mass misdiagnosis. The model would be expected to be nearly perfect to reliably diagnose people in a real world setting. Although the model can be further built up and improved, as it stands it would be insufficient for anything more than getting a general idea of whether or not an individual has COVID-19.

	Precision	Recall	F1-Score	Support
COVID-19	0.86	0.88	0.87	49%
non-COVID-19	0.88	0.86	0.87	51%

Table 1

Accuracy	0.8713
Sensitivity	0.8800
Specificity	0.8627

Table 2

Adjusting the layers of the NN would help the accuracy to an extent but the dataset used has a significantly bigger impact on these results. The biggest caveat and take away from this experiment is that, for NNs to work effectively they must have a sufficient amount of training data. If that kind of data can be provided then a NN can improve and learn overtime by tweaking their internal layers to better understand what patterns to expect. In our scenario the more X-ray images of patients with COVID-19, the more accurate our NN will be at recognizing patterns that correlate with having COVID-19. Unfortunately, X-ray datasets are limited due to the recency of this disease. As time passes it can be expected that similar technologies will only improve and be able to build off of each other in hopes of a near perfect classifier tool.

Summary

By the end of the project, a binary classifier was made to give an estimate of whether or not an individual is inflicted with COVID-19 using chest X-ray images. The estimates were made to be about 87.13% accurate. The classifier analyzes patterns that are mined using a convolutional neural network. The neural network was built using Tensorflow and was trained on a cleaned dataset of chest X-rays from patients both with and without the disease. It was discovered that by tuning the layers of the network, it could be made to be very accurate but the size of the dataset proved to be a major bottleneck. However, this only leaves hope for the future; as more data is collected and more research is done, the world as a collective will only make progress towards defeating this pandemic.

References

1. Joseph, Saumya. “Abbott Launches COVID-19 Antibody Test, Plans 20 Million Tests per Month by June.” Reuters, Thomson Reuters, 15 Apr. 2020, www.reuters.com/article/us-health-coronavirus-abbott/abbott-launches-antibody-test-for-coronavirus-plans-to-deliver-20-million-tests-by-june-idUSKCN21X21E.

2. Irfan, Umair. “The Case for Ending the Covid-19 Pandemic with Mass Testing.” Vox, Vox, 13 Apr. 2020, www.vox.com/2020/4/13/21215133/coronavirus-testing-covid-19-tests-screening.

3. Bachir. “COVID-19 Chest X-Ray Dataset.” Kaggle, 15 May 2020, www.kaggle.com/bachrr/covid-chest-xray.

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

5. Rosebrock, Adrian. “Detecting COVID-19 in X-Ray Images with Keras, TensorFlow, and Deep Learning.” PyImageSearch, 18 Apr. 2020, www.pyimagesearch.com/2020/03/16/detecting-covid-19-in-x-ray-images-with-keras-tensorflow-and-deep-learning/.

6. Sagar, Abhinav. “Deep Learning for Detecting Pneumonia from X-Ray Images.” Medium, Towards Data Science, 26 Nov. 2019, towardsdatascience.com/deep-learning-for-detecting-pneumonia-from-x-ray-images-fc9a3d9fdba8.