Covid CNN Diagnosis

Xray image recognition

## presented by

Logan Bell

X-ray of a person's chest

Description automatically generatedA chest x-ray of a person

Description automatically generated

**Covid** X-ray Left | **Normal** X-ray Right

1. Goal

I set out with the goal of demonstrating that it Is possible to diagnose covid through image recognition via CNNs, or convolutional neural networks. This is simply a proof of concept, seeing that my dataset contained many issues, and my resources were limited to a Google Collab notebook. With more time to train my model and more GPU compute power, I could use a larger dataset and get much more accurate results.

Like most people, I’ve lost multiple people very close to me to Cancer, and I think Machine Learning is underutilized in the medical industry. One person’s death was partially due to doctors not recognizing cancer in an Xray, and the diseased area was allowed to grow until it was too late. I want to start with Covid and move onto other illnesses, and train models to recognize and diagnose issues with the patient.

A diagram of a diagram of a variety of cubes

Description automatically generated

CNN model example illustration

1. Dataset

The dataset I used for this model is called the *COVID-19 Radiography Database* on Kaggle from user TAWSIFUR RAHMAN. The dataset is comparatively small in terms of the number of images and the size of the images. For large shapes like stop signs and street signs, 299x299 pixels would be plenty. But when trying to diagnose covid based on X-rays, more lung detail can be very helpful in noticing issues.

My dataset contains roughly 7000 covid, 20000 normal, and 2800 pneumonia images of lung X-rays. Due to the large disparity in the number of images I decided to only use ~3000 of each. This sounds counterintuitive, but the general idea is that the model will not resort to randomly guessing normal images more often to get a higher statistical likelihood of being correct.

In the future I might add more data to each category to increase the information available to my model. There are other kinds of lung illnesses that could also be added to my model to improve prediction accuracy.

1. Accuracy

As I glossed over previously, Covid and Pneumonia happen to look slightly similar, and Covid can be noticed in an X-ray via spotting opaque zones in the image.

When testing my model on a holdout dataset, it achieved ~90% overall accuracy, which I would not feel comfortable using for actual medical diagnoses, but I personally doubt most doctors would have a much higher rate of accuracy while looking solely at patient X-rays.

Accuracy in the real world might not be so accurate, because X-rays would hopefully be similar to my training data, or at least have similarly few obstructions in front of the lungs.

Assuming any X-rays were used and I finish using the masking process, any lungs would also need to be similarly cut-out.

I wanted to make a chart to show the ROC curve for each variable.

A graph of a function

Description automatically generated with medium confidence

The model’s learning over epochs.

This interesting graph shows the true V.S. the false positives, where a perfect graph would be bent towards the top left corner, and a bad graph would be closer to the dotted line representing being more random.

One worry that plagued my throughout the process of training my model was the nagging suspicion that it was only memorizing the data and not really learning how to actually predict Covid or Pneumonia in any way.

## Prediction accuracy spread:

You can see here that my model was less likely to pick covid, and more likely to think it was pneumonia or normal lungs when compared to the actual ratios.

This is unfortunate, and I would like to make this go the other way, because we want more false negatives than positives for health safety. If someone is told they don’t have covid when they do, that is more harmful than telling someone they do have covid when they don’t.A graph of blue and orange bars

Description automatically generated

This Demonstrates the amount of predictions the model made compared to the actual number of accurate.

I would want to somehow change this so that covid is more likely to be a false positive than false negative, and normal is more shied away from.

While it’s hard to say whether this model is ‘good’ based purely on accuracy during training, I wanted to illustrate the performance increase over time while overgoing multiple training iterations.

I feel fairly confident with these results being ~90% accurate, and I think if I went any further the model would begin memorizing the data and performance would suffer in the holdout dataset as a result.

A graph of a graph with blue and orange lines

Description automatically generated

Training Accuracy compared to Validation Accuracy over time

1. Ethical Implications

The dataset I used is public, and this might be resolved by looking closer at the description, but we can’t be sure the images were taken with the permission of all patients. It would be problematic if the patients did not know their images were being spread publicly online, but then again, that’s more of the creator’s issue and not ours. I believe that because we are not generating or spreading those images in any way, we should be okay with only using them for training diagnosis models. I don’t think generative models would have an issue, but this kind is even less problematic.

I would feel more comfortable with this model being run locally on the device, because it could be problematic if personal information was sent over the internet to company server. Because none of the data would be seen by us, all the health data would be much more secure.

1. Python Notebooks

This is the link to my Google Colabnotebook where I trained a CNN on the dataset:

[**https://colab.research.google.com/drive/1z0Ghjd02-CQhYiL\_3\_Eg17h2KMTSzuDc?usp=sharing**](https://colab.research.google.com/drive/1z0Ghjd02-CQhYiL_3_Eg17h2KMTSzuDc?usp=sharing)

And here is the dataset:  
<https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/>