Classification of COVID CT Scans

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the roll out of the vaccine, and various improvements to testing accuracy and speed, COVID-19 continues to wreak havoc in various low-income and under-developed countries. Thus, there exists the need for faster, more accurate methods of testing. We developed two neural network models for the diagnosis of COVID-19 from a lung CT scan in the hopes of resolving the challenges that currently oppose traditional testing methods.

1 Background and Problem Summary

Although the Novel Coronavirus (COVID-19) pandemic reached its peak in January of 2021[1], hundreds of thousands of new cases continue to ravage the global population. The current gold standard of treatment, the Reverse Transcriptase Polymerase Chain Reaction (RT-PCR), is both time consuming and subject to low sensitivity[1,2]. Mass testing services tend to pool samples for improved efficiency, and while this can be effective in speeding up the process, retesting is later required for any presumptive positives, and pooling done by hand can be vulnerable to human error in pipetting[2]. Furthermore, the turnaround time for an individual test is less than ideal for the confirmation of a disease as contagious as COVID-19. For presumptive cases, valuable resources are allocated towards the isolation and prevention of transmission of the disease while health care professionals await the results of the test. Due to its high sensitivity, propensity to detect not only the presence, but also the severity of COVID-19, and high availability, Computed Tomography (CT) plays a significant role in the diagnosis and management of COVID-19.

Researchers in 2021 compiled the world's largest publicly available database of COVID-19 CT scans[3]. It includes not only images of COVID-19 scans, but also scans of healthy lungs, and scans of Community Acquired Pneumonia (CAP) cases. We had two goals in mind for this data set. Firstly, we wanted to build a binary classification model which could identify whether a lung scan came from a COVID-19 patient or a healthy patient. Secondly, we wanted to expand our model to a multi-class classification problem, where identifying healthy, COVID-19, and CAP scans with high accuracy was the goal.

1.1 Data Set

We downloaded the "Large COVID-19 CT scan slice dataset" from Kaggle, available at

`www.kaggle.com/datasets/maedemaftouni/large-covid19-ct-slice-dataset`

The data set includes 7,593 COVID-19 images from 466 patients, 6,893 normal images from 604 patients, and 2,618 CAP images from 60 patients. The image data has an input shape of (224, 224, 3), output shape of (7, 7, 512), and 14714688 parameters. The database was curated from 7 public data sets listed in the references section.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Methods

2.1 Binary Classification of COVID-19 and Healthy Lung Scans

We adopted the VGG-16 neural network to train our machine learning model. First, we classified the image data into training and validation batches. Our pre-trained model for this task is VGG16, which is also the first layer of our deep neural network model. The second layer is the flatten layer, which reduces the dimension of outputs from VGG16 and lets them be ready for fully connected layers. The third layer is a fully connected layer with output dimension of 1024, 25691136 parameters, and relu as activation. The fourth layer is a drop-out layer with a drop-out ratio of 0.5. The fifth and sixth layer are another two fully connected layers with 256 and 64 output dimensions, 262400 and 16448 parameters, and relu as activation. The seventh layer is a fully connected layer, which is also the output layer and the classification layer with an output dimension of 1, 65 parameters, and sigmoid as activation. We used Adam as the optimizer and binary cross entropy as the loss function. We chose a learning rate of 0.0005 and a batch size of 10.

2.2 Multi-Class Classification of COVID-19, CAP, and Healthy Lung Scans

In this task we trained two deep neural network models to classify an X-ray image into one of the following classes: normal, COVID-19, and CAP. We chose VGG16 as a pre-trained model, which is also the first layer of our first deep neural network. It has the input shape of (224, 224, 3) and the output shape of (7, 7, 512) and 14714688 parameters. Second layer is the flatten layer, which reduces the dimension of outputs from VGG16 and lets them be ready for the fully connected layers. The third layer is a fully connected layer with 4096 output dimension, 102764544 parameters, and relu as activation. The fourth layer is a drop-out layer with a drop-out ratio of 0.5. The fifth layer is a fully connected layer with 1024 output dimension, 4195328 parameters, and relu as activation. The sixth layer is the other drop-out layer with a drop-out ratio of 0.5. The seventh and eighth layer are fully connected layers with output dimension of 256 and 64, 262400 and 16448 parameters, and relu as activation. The ninth layer is a fully connected layer, which is also the output layer and the classification layer with output dimension of 4, 260 parameters, and softmax as activation. We used Adam as the optimizer and categorical cross entropy as the loss function. We also tested the multi-class model using Inception V3 as the base, with the same layers mentioned above following it, as well as the same optimizer and loss function. We chose a learning rate of 0.0005 and a batch size of 10 for both versions of the multi-class model.

2.3 Training, Testing, and Validation Sets

For both the binary and multi-class tasks we first trained the model on a small data set of images to test the viability of the model. For the binary classification, this meant an initial training set of 160 images, and an initial validation set of 40 images, each split in half for COVID-19 and healthy lung scans. This demo model was trained for 50 epochs, and we evaluated our results before moving onto training and testing on the entire data set. For the multi-class task we had an initial training set of 240 images, and an initial validation set of 60 images, each composed of a third COVID-19, healthy, and CAP scans. Due to limited time, as well as rejected requests from BU Shared Computing Cluster (SCC) for additional resources, we could only run the training for 10 epochs with the entire data set. Due to the less-than-ideal circumstances, we have decided to not include the results we obtained with the full data set in this report; however, given more time we would have continued testing with more epochs, as well as different batch sizes, learning rates, and architectures.

2.4 Visualization of Results

Results visualization included the generation of plots for training loss and accuracy, as well as validation loss and accuracy. Final test accuracy and loss were calculated, and our model's classification was then visualized via t-distributed stochastic embedding (t-SNE) plots. The t-SNE plot was chosen as our visualization technique of choice due to its suitability for the visualization of high dimensional data sets in two dimensions[4]. T-SNE plots are included in the results section, while the specific loss and accuracy graphs for each model can be found in the appendix.

3 Results and Analysis

3.1 Training and Validation Results

For task 1, binary classification of healthy and COVID-19 scans, the average training accuracy of our model was above 0.9, and validation accuracy was 0.90 after 50 epochs. Such a pattern implies that our model is not over-fitting or under-fitting the data set, and is discerning a general pattern based on the training data. The training loss and validation loss curves implied similar results, as they were descending across 50 epochs. The training loss was around 0.25 and the validation loss was within 0.25 to 0.30 near epoch 50. The drastic drop in training loss on the early stage of epochs could be caused by the multiple back propagations on a relatively small batch of input data, given that the training loss is calculated each time the epoch terminates.

For task 2, multi-class classification of COVID-19, healthy, and CAP scans, the validation accuracy curve and the training accuracy curve for the VGG16 model were both ascending during training. However, the training accuracy was significantly higher than the validation accuracy, with training accuracy being around 0.9 and validation accuracy hovering between 0.6 to 0.7 after all epochs. Such a pattern implies that our model is able to model the training data well and also getting better to fit the new input data. However, the validation accuracy curve began to increase at around epoch 40. This could mean that our model is over-fitting, or the size of the demo data set we fed to our model was too small. Graphs for training and validation loss also showed similar patterns, as validation loss fluctuated wildly, and even started to increase around epoch 40. This again implies over-fitting, and the model's inability to accurately classify new data.

For task 2, model 2 (Inception V3), the training accuracy curve and the validation accuracy curve were both ascending after all epochs. The validation accuracy remained around 0.8 while the training accuracy was above 0.9 at epoch 100. We are able to expect the training accuracy to keep increasing to 1.0, meaning that the network is close to perfectly model the training data set. The model also displays its capability to fit a new data set, and thus make better predictions even if the model displays a certain degree of over-fitting. Compared to VGG16, the inception network suffered less from over-fitting on the same input data-set. Similarly to VGG16 however, validation loss began to fluctuate wildly and rise at around epoch 40. This could be signaling that our model was starting to over-fit at epoch 40, or that the input data set was too small. Despite this, the actual value of the validation loss for Inception V3 was still significantly lower than that of VGG16, reading approximately 0.8 and 1.2 respectively.

3.2 Test Results

The test loss for our task 1 was 0.115 and the test accuracy was 1.0. Thus of all 60 test cases, our model accurately predicted all of them, indicating that our model learned well from the training data set and also fitted the new data well.

The test loss for our task 2 VGG16 network was 1.26 and the test accuracy was around 0.717. Thus of all 60 test cases, the model accurately predicted 43 of them. The test loss is relatively high due to what is apparently an over-fitting problem. The major cause for the over-fitting problem could be attributed to that we had a limited training data set, and the model has learned noise into its prediction.

The test loss for our task 2 model 2 was 0.036 and the test accuracy was 1.0. Our second model based on the inception network is able to make better predictions on test data set with a much lower loss. The major reason could be it doesn't have as much of an over-fitting problem compared to model one. Given the same extra layers, drop rate, learning rate, and label smoothing applied to both models, the main difference between the two models lies in the distinction between Inception Network and the VGG16 Network. Inception Network includes inception modules that consist of a 1 by 1 filter followed by convolution layers with different filter sizes applied simultaneously. This allows the network to learn from more complicated features. Given that our data set has overall 235,996,451 parameters, Inception network's ability to process more parameters without over-fitting the new data as drastically could have led to the significantly higher test accuracy. However, this does come with the caveat that if we perhaps had more time to tune the VGG16 model, we could have seen much higher test accuracy.

3.2.1 Visualization via t-SNE Plots

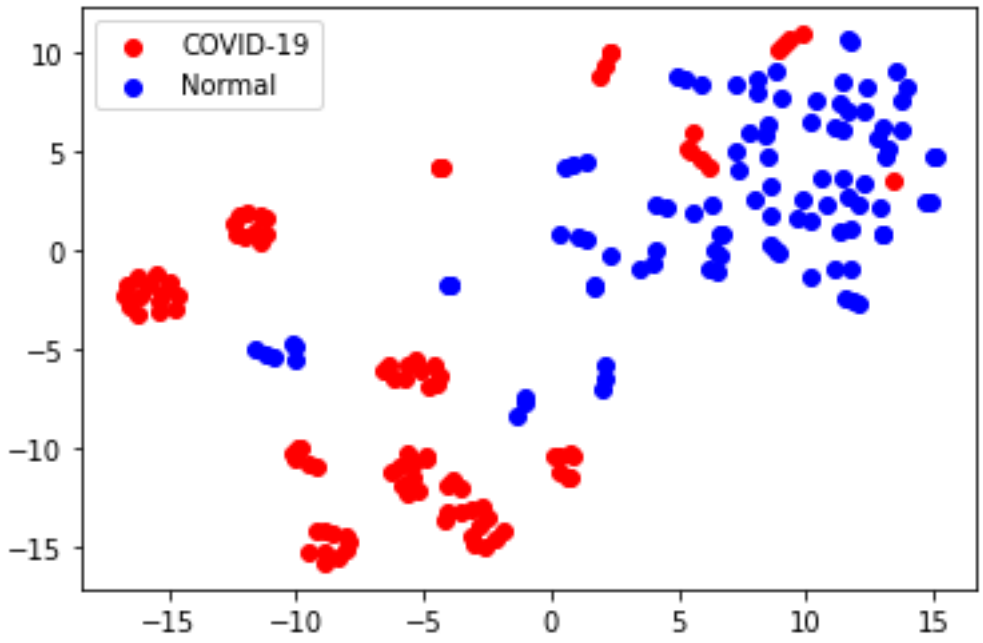

Figure 1: Task 1 binary classification t-SNE plot

Although there are a few outliers, our model was able to classify the COVID-19 and healthy lung scans into distinct clusters.

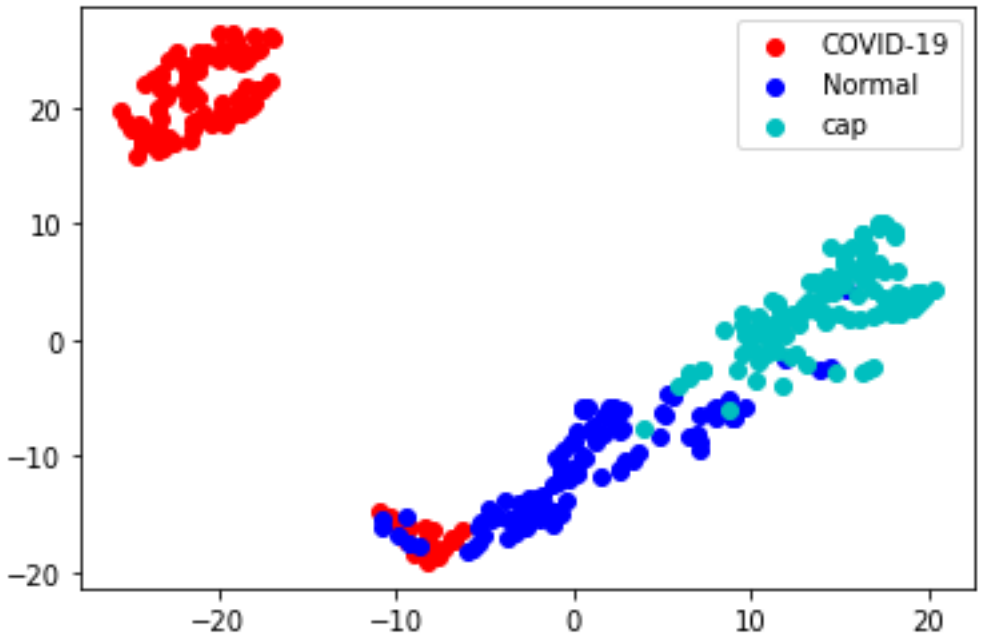

Figure 2: Task 2 multi-class classification via VGG16 t-SNE plot

The VGG16 model for Task 2 was able to separate COVID-19 cases from the rest with some duplication between the normal cases. However, our model could not separate normal and CAP CT scans into distinct clusters.

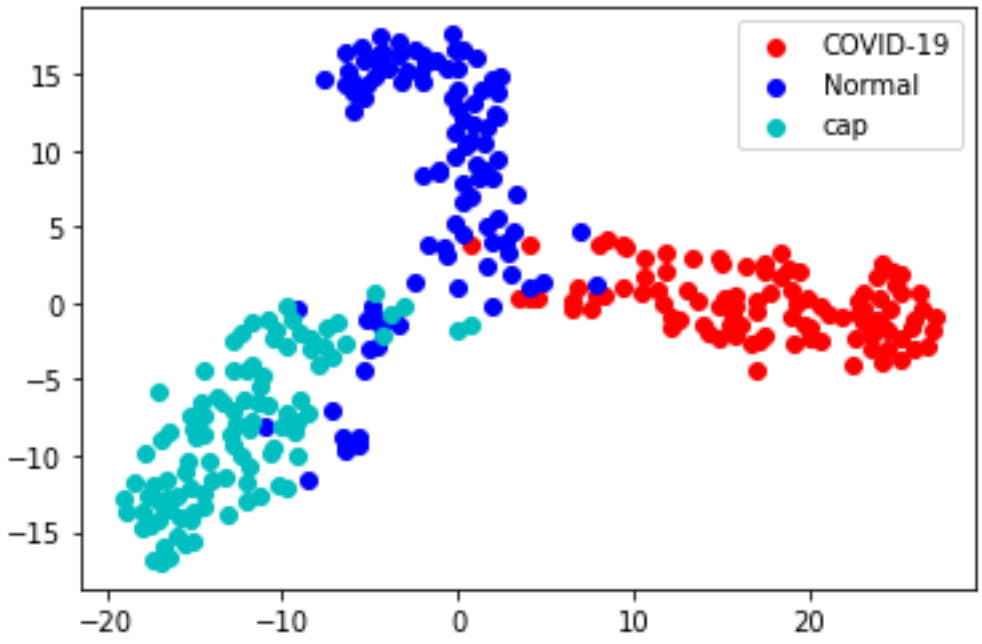

Figure 3: Task 2 multi-class classification via Inception V3 t-SNE plot

The Inception model was able to cleanly separate the images into 3 different clusters for COVID-19, normal, and CAP cases with minor overlap.

4 Conclusion

Based on our test results, the model based on inception was best at predicting if a new patient had COVID-19 or Pneumonia cases based on his/her CT scan. The model was able to achieve high accuracy with low loss on its prediction. Our model 1, trained based on the VGG16 network, was also able to identify CT scans of patients with COVID-19 from those who're normal, but not for patients who have Pneumonia. The inception network performed extremely well in handling data sets with a huge amount of parameters as well as training models in an efficient amount of time. The network suffered less from the over-fitting problem and was also able to model training data. Overall, deep neural networks were reliable when running small batches of data over numerous epochs, though our work was limited in scope due to time and resources. Possible directions for future research include training and testing the model on the full data set, as well as quantification of the severity of disease. Furthermore, it is crucial to understand that there are numerous other lung pathologies which present similar symptoms to COVID-19, including but not limited to: influenza, the common cold, seasonal allergies, and respiratory syncytial virus. Expanding this research to account for these other conditions while maintaining accuracy will require much larger data sets, and will be crucial to the viability of machine learning techniques in a diagnostic setting.

References

[1] Afshar, P., Heidarian, S., Enshaei, N. et al. (2021) COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci Data 8, 121.*

[2] Alireza Tahamtan & Abdollah Ardebili (2020) Real-time RT-PCR in COVID-19 detection: issues affecting the results *Expert Review of Molecular Diagnostics, 20:5, 453-454.*

[3] Maftouni, M., Law, A.C, Shen, B., Zhou, Y., Yazdi, N., and Kong, Z.J. (2021) A Robust Ensemble-Deep Learning Model for COVID-19 Diagnosis based on an Integrated CT Scan Images Database, *Proceedings of the 2021 Industrial and Systems Engineering Conference, Virtual Conference.*

[4] Wattenberg, et al. (2016) How to Use t-SNE Effectively, *Distill.*

A Appendix

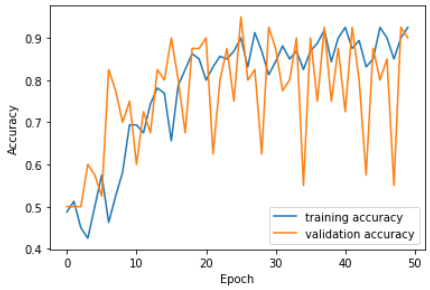

Figure 4: Task 1 binary classification training and validation accuracy

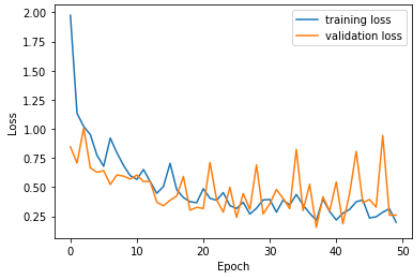

Figure 5: Task 1 binary classification training and validation loss

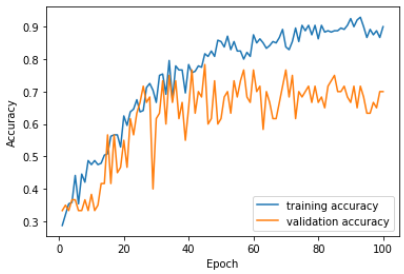

Figure 6: Task 2 VGG16 multi-class classification training and validation accuracy

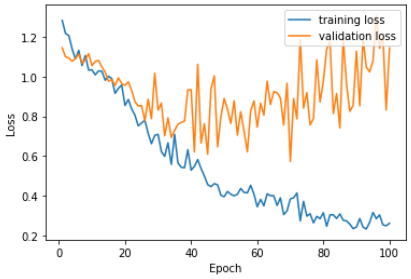

Figure 7: Task 2 VGG16 multi-class classification training and validation loss

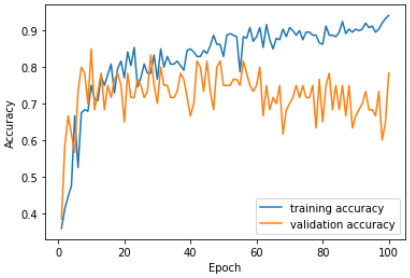

Figure 8: Task 2 Inception V3 multi-class classification training and validation accuracy

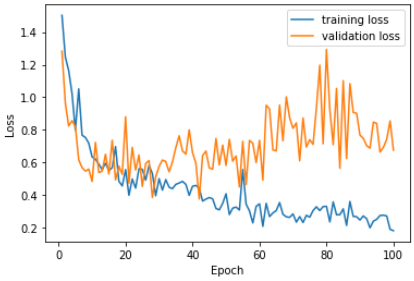

Figure 9: Task 2 Inception V3 multi-class classification training and validation loss

Code for this project can be found on github.

`https://github.com/HanlinZou/BU-CS542`