COVID-19 Pnumonia Diagnosis Using Chest X-Rays - A Machine Learning Approach

841701 Zhongyuan Zhang
824346 Yinrui Liang
102801 Yuanning Dong

June 5, 2020

1 Introduction

There is no doubt that the list of keyword in 2020 will contain the word: COVID-19. This is the short name for coronavirus disease 2019. This virus was first identified in Wuhan, China in December 2019. [Mehta et al.2020] As a science student, the first intuition is that if we can find a positive approach by what we have learned. This intuition has become the aim of this project. Researches has proven that machine learning can provide a strong help in radiology. This means that the accuracy that the machine learning techniques nowadays have reached is so high that the doctors and medical organization can utilize it for diagnosis. Not 100% trust in it, but can be a helpful reference. [Wang and Summers2012]

Therefore, the aim of this project is to build a binary classifier that can take the chest x-ray as the input and output if the patient is positive or negative in the COVID-19 diagnosis. During the project, the datasets that are obtained from a Github repository[Cohen et al.2020b] and a Kaggle dataset [Andriole2018]. They are preprocessed and pass the model. The model uses a pre-trained model as base model then it is tuned for diagnosing COVID 19. After that, the results of the model is evaluated and LIME explanation

is used to indicate the useful feature that the designed model used to distinguish the chest X-rays of a COVID patient from a normal people.

2 Literature Review

In this report, there are many papers have been read. [Cohen et al.2020a] is the main paper that gives an idea for the approach. It used CNN to predict the patients' severity in COVID. However, as we don't have any experts to determine the severity which the group in the paper author has, we can only use the model to diagnose the COVID pneumonia. In general, although the output of the model is different, we can still use their idea in preprocessing the dataset and the construction of model.

3 Background Knowledge

To understand how to diagnose COVID from X-Ray, it is required to know that the shade in X-ray normally means that there is infections in the specific part.[Belfield2010] If the shade is in the lungs, it normally means that the patient has a lung disease such as Pneumonia or lung cancer. COVID-19 is also a type of lung disease that infects the lung. Therefore, it is assumed that there will be some difference between a normal human's chest X-ray comparing to a COVID-19 patient's X-ray so that the model can distinguish them.

4 Output of the Model

For this project, the model will be considered as a binary classifier. This means that the model will only have two outputs: Normal or COVID. To make this acceptable for the model, one-hot encoder will be used.

4.1 One-Hot Encoder

One-hot text vectorization is a method that is normally used for the labels. It turns the labels into an array of either 1 or 0. The steps of doing this

is build a table that contains all the genres first. In this case the table is expressed in the following order: COVID, Normal. Then, the encoder will have 1 as the label belongs to the instance and 0 means no. For example, if the patient's X-Ray is positive for COVID, the encoded array by one-hot encoder will be [1,0].

5 Data

There are two categories of data used in the model: the chest X-rays from a normal person and chest X-rays from a COVID-19 patient, the samples of these two categories are shown in figure 1. Among all the data that has been provided, only one view of the X-rays can be chosen so that the model can find the specific pattern of the samples. In this project, the PA view is chosen.

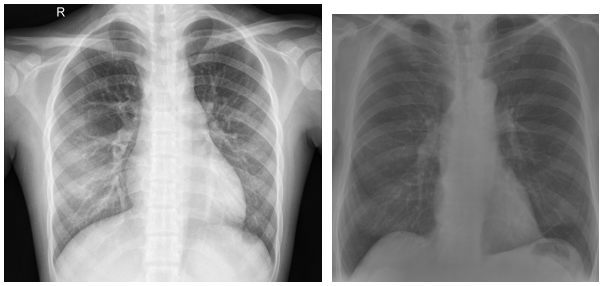

Figure 1: L: Sample Chest X-Ray of a Normal Person, R: Sample Chest X-Ray of a COVID-19 Patient

As it can be seen above, the normal patient's lung appears to be more transparent than that of the COVID patient's. This is because the coronavirus has infected the COVID 19 patient. This set of comparison has further verify the reliability of the assumption that was made in section 2.

6 Pre-Processing the Data

To make the images become acceptable for the model, the images will be resized to 224×224 pixels. The values will be normalized into a range of 255.0. The resize is done by a centre crop. Also the colour space of the picture is converted by the command in cv2 library called cvtcolor. For here,

we use BGR2RGB so that the colour ordering is changed which allows the Convolutional Neural Network(CNN) to process.

7 Model selection

VGG-16 network is a simple-structured Convolutional Neural Network which only consists of max pooling layers (2×2) with a stride of 2 and convolution layers that just use 3×3 filters with stride of 1 and same padding. It contains 16 layers and has a total of about 138 million parameters which make it a pretty large network. But what makes it appealing is the simplicity of its architecture. Its architecture is quite uniform whose pattern is always a few convolution layers followed by a pooling layer. A more deeper network Resnet 50 is used to train the model in this project. In theory as a neural network becomes deeper it is supposed to perform better on training set. But empirically, for plain network, the training error will tend to decrease after a while and then it will go back up as the number of layers increases. This is due to vanishing and exploding gradient types of problems. He et al. 2015 introduces skip connections which enable researchers to train very deep networks of over 100 layers. The strength of skip connections is that they can allow the model to learn an identity function which ensures that the higher layer will perform at least as good as the lower layer.[Dwivedi2019] Residual blocks are convolution layers with skip connection as figure 2 shows, and Resnet 50 are stacked with these residual blocks to make 50 layers. Resnet Paper shows that for some specific image recognition problems Residual 50 can perform better than VGG-16. A pre-trained Resnet 50 is used in this project due to the small volume of the training set. The output of the pre-trained model is passed to a dense layer which is used for the further fine-tuning so that he model can be used to diagnose COVID. Dropout layer is included to avoid the problem of over-fitting. Finally, a softmax activation is used so that the layer will have one choice between normal or COVID.

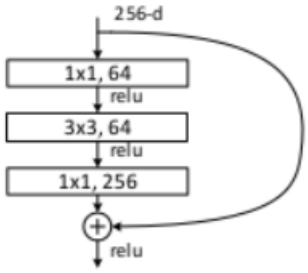

Figure 2: Residual block in Resnet 50

8 Result

In this project, 40% of the datasets are used for testing the accuracy of the model. The VGG model reaches an training accuracy of 94.25%. Meanwhile, the training accuracy of ResNet50V2 reaches 100%. The training accuracy are used to compare the performance of the two models due to the small volume of dataset. If additional test dataset is required, the remaining size will be too small for training which cannot make the model be trained as much as expected. Hence, ResNet50V2 is considered as the better model. As it can be shown below: the model reaches an accuracy of 98.7%. The pre-trained model has made an important contribution in increasing the accuracy significantly. The follow image has shown the learning curve of this model.

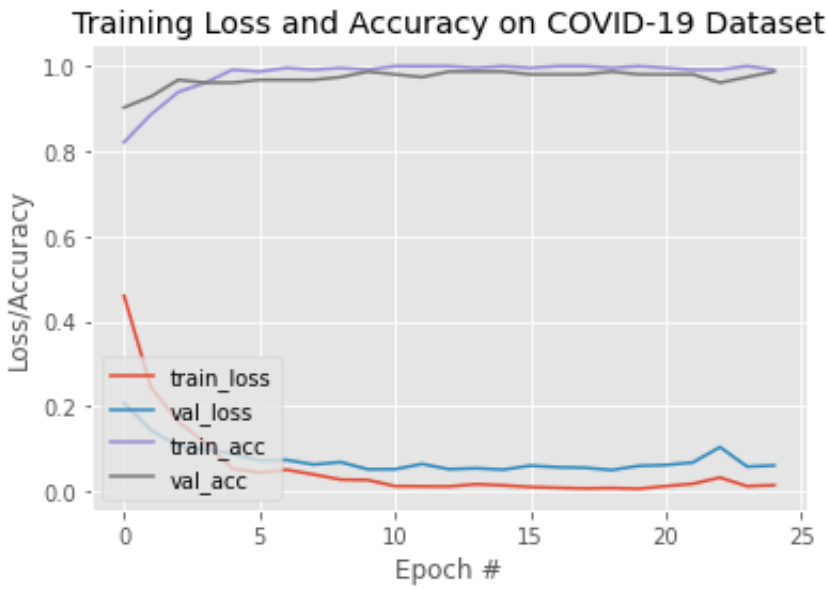

Figure 3: Learning Curve of the Model

As it can be seen above, the model has converged, this means that more epochs will not further improve the model. The model has reached the best performance based on the current circumstance. Besides accuracy, it is considered that the model should be evaluated with recalls as the metrics. Recall is used to show how many actual positive instances are predicted as positive by the model. Considering that the disease is extremely contagious, recall has to have the largest weights among all the metrics. From the figure below, it can be seen that the recall for COVID class reaches 97%.

	precision	recall	f1-score	support
COVID	1.00	0.97	0.99	76
Normal	0.97	1.00	0.99	78
accuracy			0.99	154
macro avg	0.99	0.99	0.99	154
weighted avg	0.99	0.99	0.99	154

Figure 4: Classification Report for the Model

Although this is high value, it is not good enough in the realistic circum-

stance. The concution matrix is shown below:

$$\begin{bmatrix} 74 & 2 \\ 0 & 78 \end{bmatrix} \tag{1}$$

The confusion matrix has shown that there are two patients that are actually positive in COVID 19 but the model predict them as normal. If this model is really used in the hospital. These two patients will be released by the hospital and due to the high infectious of COVID, the consequences are very serious.

9 LIME Explanation

Local Interpretable Model-Agnostic Explanation, also known as LIME explanation, is a method for explaining the predictions of any classifier. The explanation indicates the features that the model used as the most important feature to determine the specific label. The following two figures shows the explanation for a normal chest X-ray and a COVID 19 chest X-ray respectively.

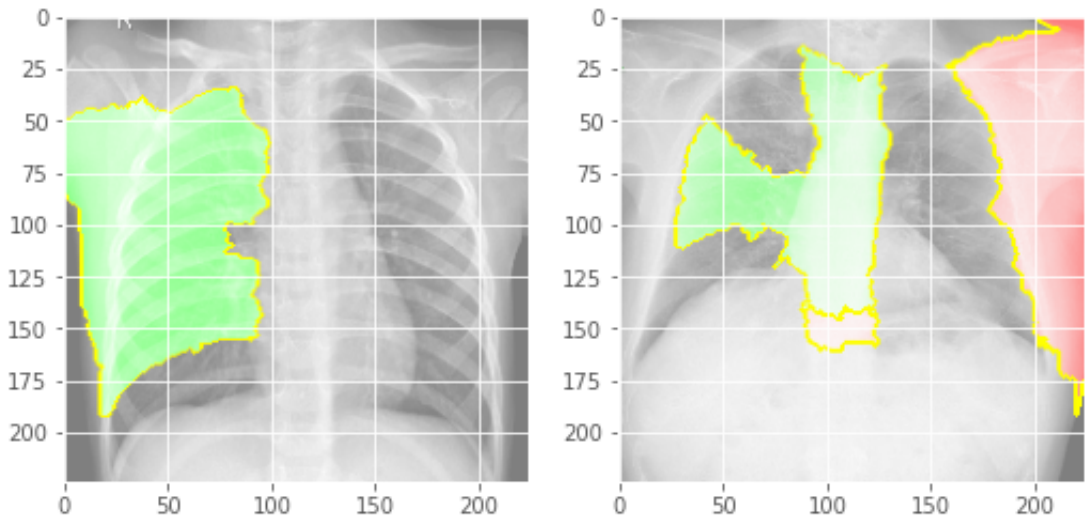

Figure 5: L: LIME Explanation of a Normal Chest X-Ray, R: LIME Explanation of a COVID-19 Chest X-Ray

By Comparing these two above, the green area shows the top features that make the model decide the label of the instance. As it can be seen above, the most important features are all in the section in the left lung. This means that the model uses the difference in the left lung to determine the label. This behaviour matches with the assumption made before that the COVID-19 patient will have infections in the lung so that their X-rays will be less transparent than the normal X-Ray. The red area indicates the top features that make model consider that the instance is the opposite. In this case, the COVID-19 has the red area in the right because there is a significant dark area. The high contrast between the lung and the dark area has made model consider the x-ray has a transparent area there.

10 Discussion

Although this model has reach a high accuracy based on the given dataset, we cannot state that this model performs well enough to practically use the model. This is due to different reasons.

1. The size of the dataset is too small. Although people are paying effort in collecting the dataset, there are still only few datasets for the model to learn comparing to the normally used dataset size which is more than thousands and even millions. Therefore, even though we used 40% of the dataset as the test dataset, there are only 154 cases in total. Considering that there are 6.29 millions of people confirmed to have COVID, 154 is not a meaningful number to prove the validity of the model in a practical circumstance. To improve this, more data are required to collect and this will need all of the people around the world to help with it .

2. The model has only proved that it can distinguish COVID X-ray from normal X-ray in a theoretical circumstance. However, it is not tested that if the model will be available to distinguish COVID from a normal pneumonia. Normal x-ray have significant difference between COVID-19 patients because normal people does not have any infections in the lung but regular pneumonia patients also have infections in the lung. Further researches for checking if the infections due to COVID-19 are different to that of regular pneumonia will be required to be done. Also, this requires a larger dataset size so that the practical validity can be verified.

3. As it can be seen in the metadata from the Github repository [Cohen et al.2020b], there are more informations about each patient. For example, whether the patient needs to be sent to the ICU, and whether the patient needs supplemental. These information is assumed that they can also be learned by the model so that the model can predict these labels based on the X-ray.

4. In this project, the pre-trained ResNetV2 is used as the base model. However, there are many different kinds of pre-trained model. For example, DenseNet from TorchXRay Vision Library has proven to be a efficient model in predicting pneumonia.[Cohen et al.2020a] Therefore, it is also worth to try the model and the performance of it can be compared to the current model.

11 Conclusion

In this project, the aim is to build a machine learning model that is able to diagnose COVID-19 by the chest X-rays. To achieve this, images are preprocessed and pre-trained ResNetV2 model is used. In the end, the model reaches an accuracy of 97%. Although this is a high value but the limitation of dataset size cannot prove the validity of the model in the practical circumstance. In the end, further actions that we can do to the model is discussed.

References

[Andriole2018] Katherine P. Andriole. 2018. Rsna pneumonia detection challenge.

[Belfield2010] Jane Belfield. 2010. Using gagne's theory to teach chest x-ray interpretation. *The clinical teacher*, 7(1):5–8.

[Cohen et al.2020a] Joseph Paul Cohen, Lan Dao, Paul Morrison, Karsten Roth, Yoshua Bengio, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Marzyeh Ghassemi, Haifang Li, and Tim Q Duong. 2020a. Predicting covid-19 pneumonia severity on chest x-ray with deep learning.

[Cohen et al.2020b] Joseph Paul Cohen, Paul Morrison, and Lan Dao. 2020b. Covid-19 image data collection. *arXiv 2003.11597*.

[Dwivedi2019] Priya Dwivedi. 2019. Understanding and coding a resnet in keras. *towards data science*.

[Mehta et al.2020] Puja Mehta, Daniel F McAuley, Michael Brown, Emilie Sanchez, Rachel S Tattersall, and Jessica J Manson. 2020. Covid-19: consider cytokine storm syndromes and immunosuppression. *The Lancet*, 395(10229):1033–1034.

[Wang and Summers2012] Shijun Wang and Ronald M. Summers. 2012. Machine learning and radiology. *Medical Image Analysis*, 16(5):933 – 951.