BABES BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMANIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

COVID-19 Detection

– MIRPR report –

Team members
Condor Mara, Mathematics Informatics, 831
Gavrilovici Eduard, Mathematics Informatics, 831
Goia Iulian Mihai, Mathematics Informatics, 831

2020

Abstract

Originating in Wuhan, China, COVID-19 is a respiratory virus that has spread quickly all over six continents and it had a huge impact on everything from the normal daily life to the global economy, affecting different aspects of public and private life, from economic and environmental fluctuations to changes that affect individuals in terms of income, education, employment. Also known as SARS-CoV-2, in the more serious cases this virus causes lung lesions and pneumonia and it is of high importance that the positive cases to be discovered as early as possible. Having such a devastating effect, the need for more diagnostic tools has emerged and based on recent studies, radiology imaging may contain important information about the virus and that is why paired with artificial intelligence algorithms it can be helpful in detecting with accuracy the presence of the virus.

Contents

Chapter 1

Introduction

1.1 What? Why? How?

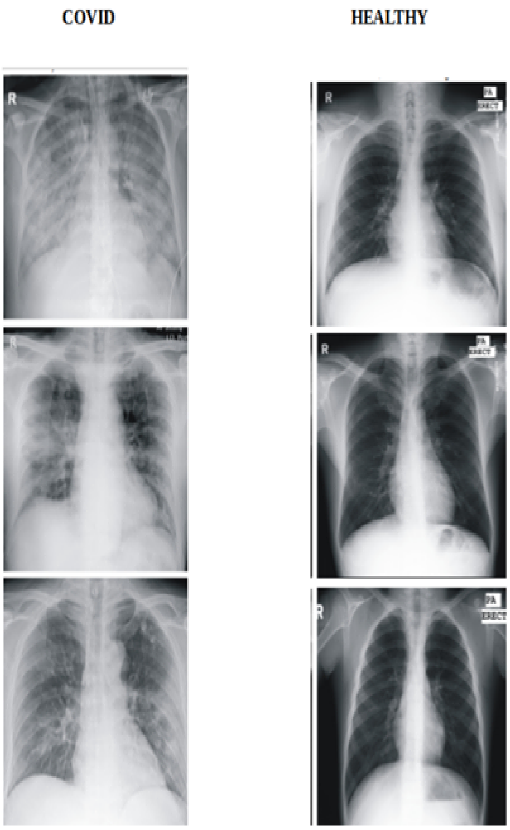
The demand for other diagnosis measures in order to detect COVID-19 has increades due to the shortage of the existing testing kits. Taking into consideration that doctors tend to use X-rays and CT scans to diagnose pneumonia and lung inflammation they can also be used to test the presence of the virus.

Chapter 2

Scientific Problem

2.1 Problem definition

The research presented in this paper advances the theory, design, and implementation of several particular models.
The main purpose of this report is to present the implementation of an intellient algorithm for solving the problem of detecting the presence of the COVID-19 virus based on a an X-ray with the best accuracy possible. This problem can only be solved by using a Convolutional Neural Network (CNN) where the neurons in one layer do not connect to all the neurons in the next layer and it rather uses a three-dimensional strucure, each set of neurons analyzing a specific region of the image. Also, they have proven to be the best method due to the self-learning ability they represent.
The report is aimed at solving the problem of processing X-ray images and classifying them as positive/negative cases.

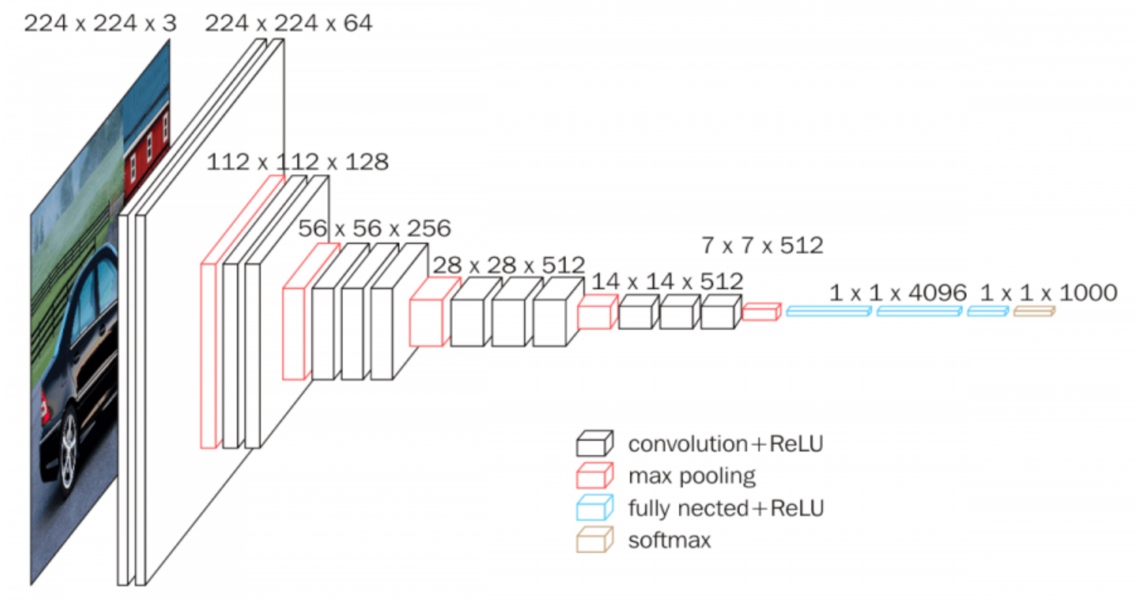


Chapter 3

Proposed Approaches

3.1 First Approach

The first approach tackeled is represented by the use of a VGG16, a convolutional neural network model with weights pre-trained on ImageNet, leaving off the fully connected layer head. All of VGG's hidden layers use ReLU, an activation function responsible for transforming the summed weighted input from the node into the activation of the node or output for that input. Given that this is a 2-class problem, we use "*binary_crossentropy*" loss rather than categorical crossentropy. We first make predictions on the testing set and grab the prediction indices and a confusion matrix was used for further statistical evaluation, to derive the accuracy, sensitivity, and specificity and print each of these metrics.



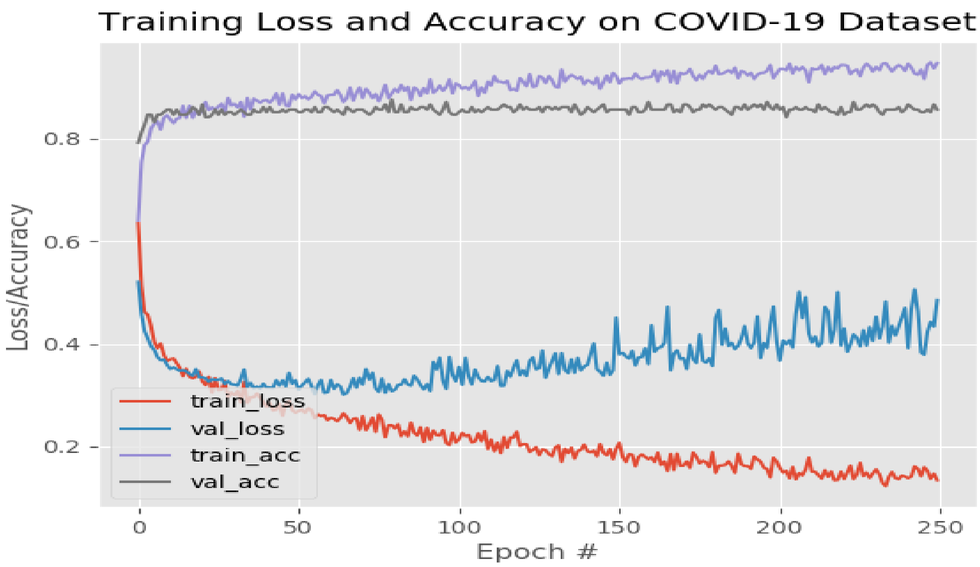
VGG-Architeture

The model was trained on a dataset constructed from two public datasets which contained X-rays, CT scans. The training data was split in two classes labeled covid and non-covid using an 80-20 split. It was trained in 250 epochs with a batch size of 32 and an initial learning rate of 0.0001, the accuracy obtained in the end being only 85 % due to the mixed types of images and not taking into account the view position DICOM attribute, which specifies the position of the radiographic view of the image relative to the imaging subject's orientation. One-hot encoding of labels was performed meaning that the data was of the following format: each encoded label consisted of a two element array with one of the elements being hot (i.e., 1) versus not (i.e., 0). In order to ensure that our model generalizes, we perform data augmentation by setting the random image rotation setting to 15 degrees clockwise or counterclockwise.

```
              precision    recall  f1-score   support

       covid       0.79      0.97      0.87       101
   non_covid       0.96      0.74      0.84       101

    accuracy                           0.86       202
   macro avg       0.88      0.86      0.85       202
weighted avg       0.88      0.86      0.85       202

[[98  3]
 [26 75]]
acc: 0.8564
sensitivity: 0.9703
specificity: 0.7426
```

97 % sensitivity and 74 % specificity was obtained implying that:

- Of patients that do have COVID-19 (i.e., true positives), the model could accurately identify them as COVID-19 positive 97 % of the time
- Of patients that do not have COVID-19 (i.e., true negatives), the model could accurately identify them as COVID-19 negative only 74 % of the time



The training history plot showing accuracy and loss curves demonstrates that the model is not overfitting despite limited COVID-19 X-ray training data used.

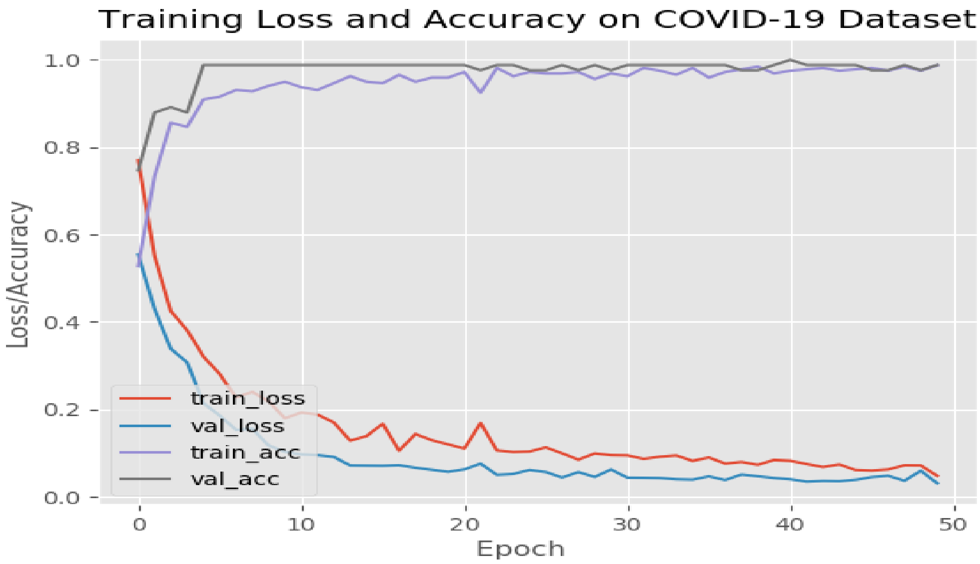
3.2 Second Approach

For the second aproach we used the same convolutional neural network model VGG16, but a different dataset was used with a different number of epochs and batches. This time the dataset was ”cleaner”, consisting only of X-ray scans with the image position being PA = Posterior/Anterior. With this aproach the accuracy obtained is 98%.

```
             precision    recall  f1-score   support

       covid       0.98      1.00      0.99        41
   non_covid       1.00      0.98      0.99        42

    accuracy                           0.99        83
   macro avg       0.99      0.99      0.99        83
weighted avg       0.99      0.99      0.99        83

[[41  0]
 [ 1 41]]
acc: 0.9880
sensitivity: 1.0000
specificity: 0.9762
```

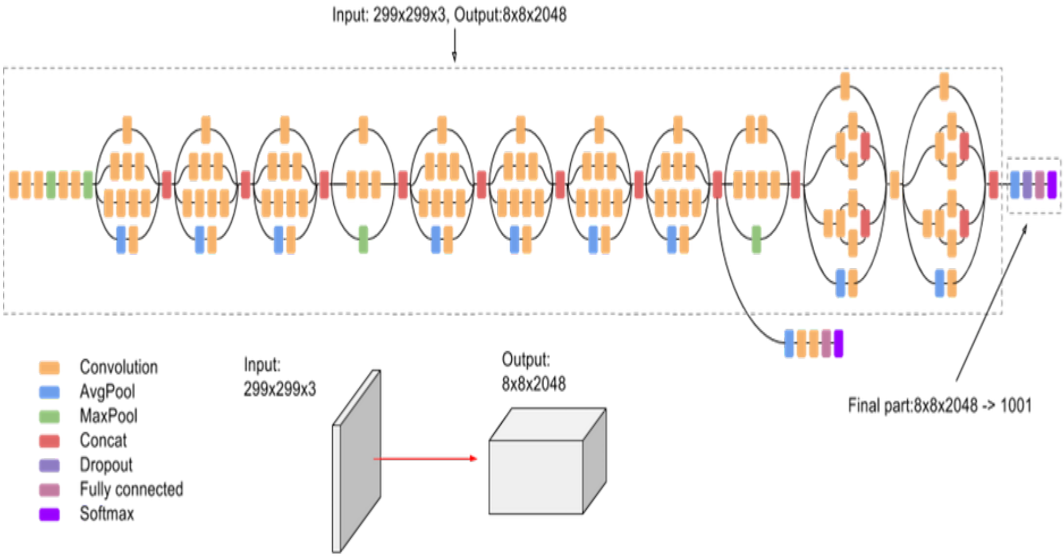
100 % sensitivity and 97 % specificity was obtained implying that:

- Of patients that do have COVID-19 (i.e., true positives), the model could accurately identify them as COVID-19 positive 100 % of the time
- Of patients that do not have COVID-19 (i.e., true negatives), the model could accurately identify them as COVID-19 negative 97 % of the time



3.3 Third Approach

For the last approach the InceptionV3 model was used. The model itself is made up of symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Batchnorm is used extensively throughout the model and applied to activation inputs. Loss is computed via Softmax. With this aproach the accuracy obtained is 97%.

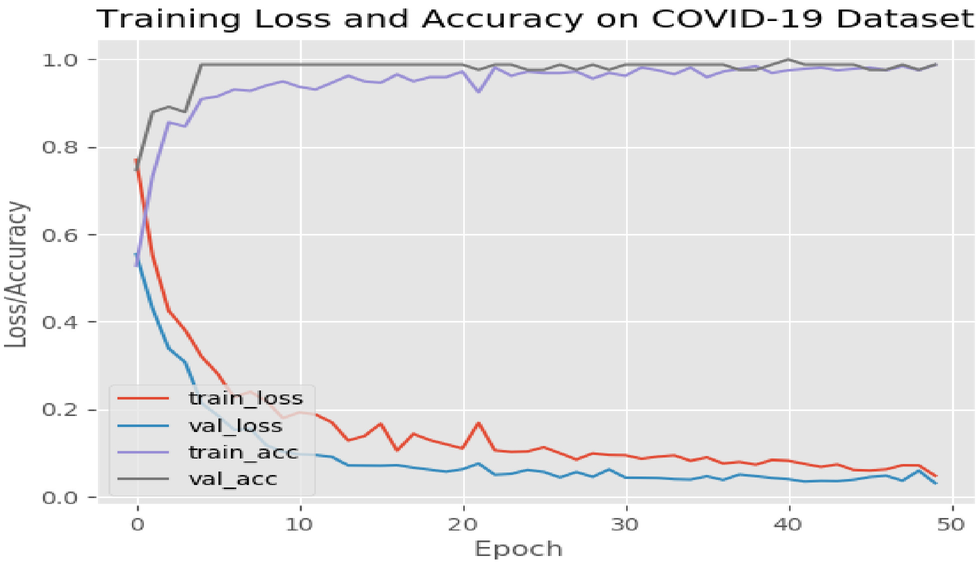


The model was trained on a dataset constructed from two public datasets which contained X-rays, CT scans. The training data was split in two classes labeled covid and non-covid using an 80-20 split. It was trained in 80 epochs with a batch size of 16 and an initial learning rate of 0.00001.

```
              precision    recall  f1-score   support

       covid       1.00      0.95      0.97        41
   non_covid       0.95      1.00      0.98        42

    accuracy                           0.98        83
   macro avg       0.98      0.98      0.98        83
weighted avg       0.98      0.98      0.98        83

[[39  2]
 [ 0 42]]
acc: 0.9759
sensitivity: 0.9512
specificity: 1.0000
```

95 % sensitivity and 100 % specificity was obtained implying that:

- Of patients that do have COVID-19 (i.e., true positives), the model could accurately identify them as COVID-19 positive 95 % of the time
- Of patients that do not have COVID-19 (i.e., true negatives), the model could accurately identify them as COVID-19 negative 100 % of the time


Training Loss and Accuracy on COVID-19 Dataset
Loss/Accuracy
Epoch
train_loss
val_loss
train_acc
val_acc
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

Chapter 4

Conclusions

- The data available is not (reliable) enough to train a COVID-19 model.
- For the COVID-19 model to be deployed in the field, it would have to go through rigorous testing by trained medical professionals, working hand-in-hand with expert deep learning practitioners. The method covered here today is certainly not such a method, and is meant for educational purposes only.
- There is no certainty that the model is actually "learning", it's possible that the model is learning patterns that are not relevant to COVID-19, and instead are just variations between the two data splits (i.e., positive versus negative COVID-19 diagnosis).
- Future (and better) COVID-19 detectors should be multi-modal. Right now the models are using only image data (i.e., X-rays) better automatic COVID-19 detectors should leverage multiple data sources not limited to just images, including patient vitals, population density, geographical location, etc. Image data by itself is typically not sufficient for these types of applications.