COVID-19 Radiography Classification with Convolutional Neural Networks

Klebert Toscano de S. Cintra

Contents

Introduction

Overview

The year 2019 will be remembered as the year the world discovered a disease that had impact on economies, politics, and changed priorities and habits for everyone, everywhere. The pandemic united the scientific community in the search for understanding of the causal agent (the coronavirus named SARS-CoV-2), the characteristics of the disease named COVID-19, its physiopathology and transmission mechanisms, diagnosis, and candidates for a cure, which is still ongoing, with several vaccines in development concomitantly and tested for efficacy and safety (as of March, 2021) with no definitive cure for the disease or immunization against the continuously mutating coronavirus.

Nevertheless, there is treatment to minimize the damage caused by the virus. The early treatment for COVID-19 patients influences the prognosis of the disease and the severeness of its symptoms, hence the importance of an early diagnosis. The complication for this is that many diseases have similar symptoms[1] and the differential diagnosis becomes quite difficult at times. The gold standard test for the presence of the virus in the organism and subsequent diagnosis is the RT-PCR, which detects the genetic material (RNA) of the virus. The

[1] Other infectious agents that cause symptoms similar to COVID-19 are Influenza A, Influenza B, Influenza A H1N1 pdm09, influenza H3N2, human para-influenza virus (HPIV), respiratory syncytial virus, rhinovirus, adenovirus, human metapneumovirus, human bocavirus, human coronavirus (HCoV), Chlamydia, Mycoplasma pneumoniae and other that can be endemic for specific regions like Dengue. For more information see this review and the list on the British Medical Journal.

test is not always available in test centers and clinics given the high demand, but a study by Vieceli et al.(2020) points to a strong predictive power of the association of leukocyte count, LDH[2] levels, and chest radiographic abnormalities for the detection of COVID-19 with the Area under the ROC curve of **0.827**, **96% sensitivity** or recall and 73.5% specificity.

[2] LDH is a predictor of inflammation in several lung diseases, and most studies indicate it to be a good predictor of severity and ICU admission.

The main radiographic findings in patients with pneumonia who tested positive for SARS-Cov-2 include:

Ground-glass opacity: signal that a substance other than air is filling the region, increasing the density. This was the most specific radiological finding among COVID-19 patients (65.6% of COVID-19 positive sample).

Lower lobe predominance: concentration of the abnormalities in the lower region of the lungs (86.2% of COVID-19 positive sample).

Bilateral involvement: both lungs present abnormalities (75.9% of COVID-19 positive sample).

Consolidation: swelling or hardening of the soft tissue as a result of being filled with liquid instead of air (51.7% of COVID-19 positive sample).

Infiltration: presence of a substance denser than air within the lung parenchyma, such as pus, blood or protein (44.8% of COVID-19 positive sample).

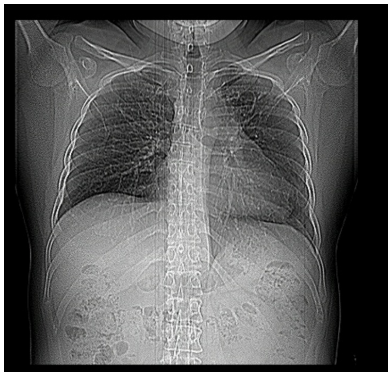


Figure 1: Chest radiography of a patient showing the signs of viral pneumonia by SARS-Cov-2.

X-ray images of the lungs are indicated when risk factors for disease progression are present (Rubin et al. 2020), but we work with the hypothesis that further investigation of images at all stages of the disease using neural networks can improve the specificity of the diagnosis and the early intervention that can change the final outcome.

The best algorithms for categorization of images into classes, like the diagnosis that corresponds to the x-ray of a patient, are based on Convolutional Neural Networks. A Neural Network is a system where the inputs are automatically transformed in a way to learn output patterns. The metaphor with the biological neural networks found in the nervous system of animals relies on the neuron, that in biological nervous systems is the functional unit. In an artificial neural network it is an abstraction consisting of a mathematical function. The values in the data go through the following stages:

1) the function takes in values as inputs in a multidimensional data structure called a *tensor*;

2) aggregates the inputs into a unique value;

3) assigns to each one of the input sources a weight that represents its relevance to make a prediction;

4) the aggregated value is passed on to an activation function, which defines the final output for that particular neuron. The output

is compared to a known correct value of the dimension the system is to predict, and the errors in the prediction are used to correct the weights of that neuron.

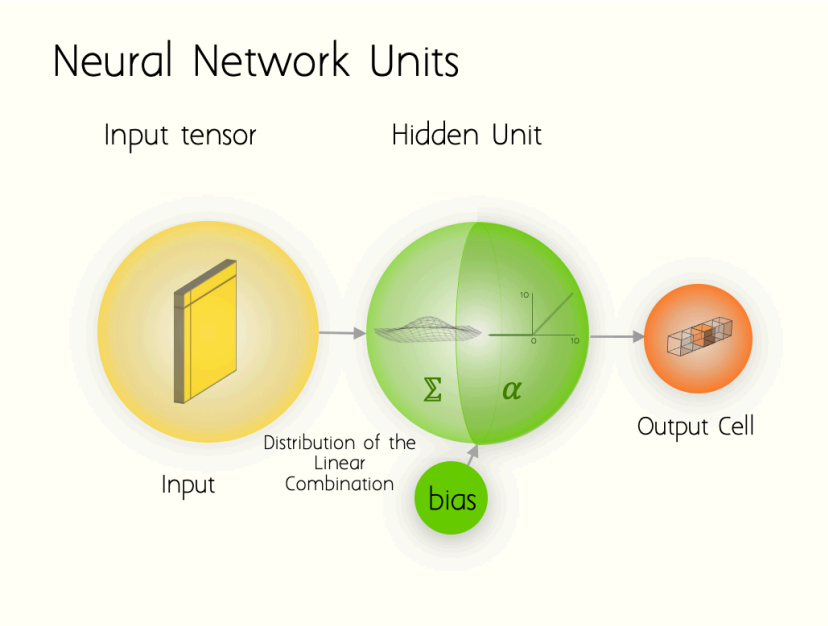


Figure 2: Illustration of an Artificial Neuron with representation of inputs, aggregation of inputs and production of an output by the activation function.

Neurons can be organized in a great variety of ways, and the process of adjusting all the parameters of the chain of functions is called *training*, borrowing a nomenclature used in cognitive psychology and learning neuroscience. The architectures proposed for this analysis has neurons that use convolutions to learn weights for the preceding layer of inputs. They are called *filters* or *kernels* and usually have small receptive fields, meaning they use a small set of values of the larger array of inputs. They are convolved across the input tensor computing the dot product to produce a 2D *activation map* as output. That map represents a *feature*, and will be passed forward to the next layer of neurons, where higher and higher levels of abstraction will be represented, in such a way that the spatial dependency is lost.

Some very desirable attributes emerge from this type of processing of data. First, there is no need to store values of the result of the interaction of each input and output like in a fully connected network. Instead, the filter, smaller than the input, will learn relevant information in fewer values. This produces a more economical algorithm, with broader applications of the parameters learned, as the layers represent levels of abstraction that can be applied to other spatial locations in the image, that can also be transferred to other data sets, and other types of output.

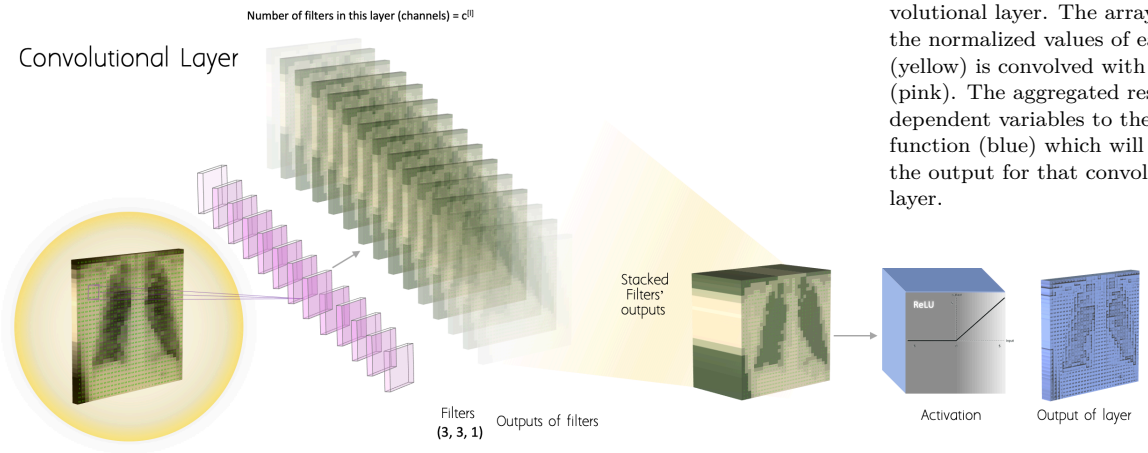


Figure 3: Representation of a convolutional layer. The array with the normalized values of each pixel (yellow) is convolved with the filters (pink). The aggregated results are the dependent variables to the activation function (blue) which will generate the output for that convolutional layer.

Comparison: Biological vs Artificial Vision

The visual system in animals has neurons that take in some information/stimuli and produce a response. Much similarly, the functions of artificial neural networks transform the data. In the case of the human vision, photons go through the eye and reach the retina, which is basically a flat layer of neurons with specific molecules that change in the presence of light, and provoke electrical variations in the voltage of the cell. This perturbation of the electrical state are like the different values you can find in the pixels of a picture. By itself, without context or relationships with other levels of abstraction, a pixel has no meaning. The same way, a neuron firing in the eye is not "vision". But when that raw particle of information is sent to the back of the brain, it reaches layers of neurons organized in the same topography as the retina, and the spatial relationships of the origin of those impulses are preserved until that point. Those neurons exchange information between them, and send signals forward to another layer of neurons which don't fire for a photon in the eye. Instead, they fire when a line of neurons in the previous layer has fired together in a specific angle. Now we have code for lines, borders, and such. The complexity of information increases at each layer up to the point where you have populations of neurons that respond to specific colors, textures, shapes, positions in the space, semantic meaning, familiar faces, and can provoke complex and systemic reactions like the fear of a spider, or catching a ball mid air.

The artificial neural networks used here also have simpler features

at the lower layers, encoding lines, curves, angles, shapes, spatial frequencies, and the goal is to make them learn higher and higher levels of abstraction up to a point where the radiography of a lung turned into a matrix of pixels can be "interpreted" as just a normal image of a healthy lung or a diagnosis for COVID-19 or other types of pneumonia. All that should happen without a formal definition of the radiographical signs mentioned before. In other words, we don't have to code an algorithm to detect lung opacity due to the presence of liquid, inflammation of the airways, or any specifics of the image that demands specialized knowledge in pneumology to be detected.

Executive Summary

The Goal of this project is to apply Convolutional Neural Networks to a dataset of chest radiographies of COVID-19 infected patients, normal subjects, and patients with other types of viral pneumoniae and test if the accuracy is improved when compared to the association of laboratory tests and traditional radiological analysis used in the aforementioned study.

The procedure proposed here includes the following steps:

1. Data acquisition. Download and partition of the data with 10% of observations for test and 90% for training of the algorithms.
2. Exploratory Data Analysis (EDA) and Visualization.
3. Methods for the analysis.
4. Modeling approach 1 - Convolutional Neural Network.
5. Modeling approach 2 - Transfer Learning - NASNet. (NASNet-Large).
6. Results - Evaluation of models and comparison.
7. Conclusion and final considerations.

Data Acquisition

The data was assembled and prepared by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia in collaboration with medical doctors. The data sources[3] are:

Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 DATABASE: SIRM COVID-19 database reports 384 COVID-19 positive radiographic images (CXR and CT) with varying resolution. Out of 384 radiographic images, 94 images are chest X-ray images and 290 images are lung CT images. This database is updated in a random manner and until 10th May 2020, there were 71 confirmed COVID-19 cases were reported in this database.

[3] Further information on the methods and data sources are available in Chowdhury et al., 2020 and Rahman et al., 2020.

Novel Corona Virus 2019 Dataset: Joseph Paul Cohen and Paul Morrison and Lan Dao have created a public database in GitHub by collecting 319 radiographic images of COVID-19, Middle East respiratory syndrome (MERS), Severe acute respiratory syndrome (SARS) and ARDS from the published articles and online resources. In this database, they have collected 250 COVID-19 positive chest X-ray images and 25 COVID-19 positive lung CT images with varying image resolutions. However, in this study, authors have considered 134 COVID-19 positive chest X-ray images, which are different from the images of the database that the authors created from different articles.

COVID-19 positive chest x-ray images from different articles: GitHub database has encouraged the authors to look into the literature and interestingly more than 1200 articles were published in less than two-months of period. Authors have observed that the GitHub database has not collected most of the X-ray and CT images rather a small number of images were in that database. Moreover, the images in SIRM and GitHub database are in random size depending on the X-ray machine resolution and the articles from which it was taken. Therefore, authors have carried out a tedious task of collecting and indexing the X-ray and CT images from all the recently publicly available articles and online sources. These articles and the radiographic images were then compared with the GitHub database to avoid duplication. Authors managed to collect 60 COVID-19 positive chest X-ray images from 43 recently published articles, which were not listed in the GitHub database and 32 positive chest x-ray images from Radiopaedia, which were not listed in the GitHub database.

COVID-19 Chest imaging at thread reader: A physician has shared 103 images for 50 different cases with varying resolution from his hospital in Spain to the Chest imaging at thread reader. Images from RSNA-Pneumonia-Detection-Challenge database along with the Chest X-ray Images database from Kaggle were used to create the normal and viral pneumonia sub-databases of 1579 and 1485 X-ray images respectively.

RSNA-Pneumonia-Detection-Challenge: In 2018, Radiology Society of North America (RSNA) organized an artificial intelligence (AI) challenge to detect pneumonia from the chest X-ray images. In this database, normal chest X-ray with no lung infection and non-COVID pneumonia images were available.

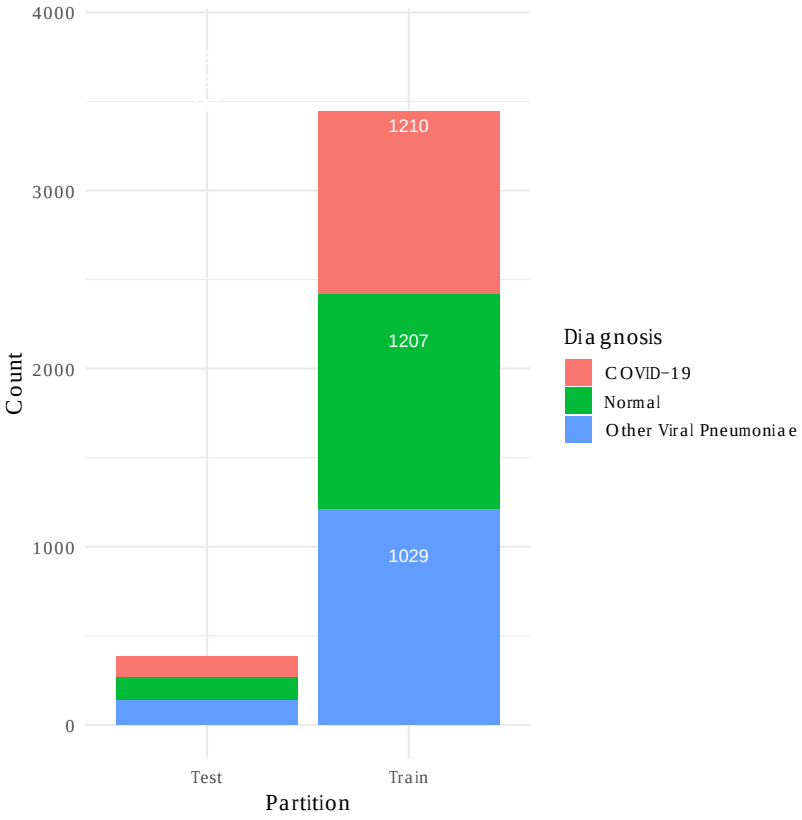
Chest X-Ray Images (pneumonia): Kaggle chest X-ray database is a very popular database, which has 5856 chest X-ray images of normal, viral and bacterial pneumonia with resolution varying from 400p to 2000p. It contains images of patients affected by pneumonia (bacterial and viral) and from normal subjects. Chest X-ray images for normal and viral pneumonia were used from this database to create

the new database.

Exploratory Data Analysis (EDA) and Visualization.

The data consists of chest X-ray images (radiographies) for COVID-19 positive cases (1200), Normal images with no infection (1341) and Viral Pneumonia (1345) images, that come in different sizes. The plot below shows the proportion of the dataset for each class, with no particular benefit for a model bias to any of the classes.

The assignment of images to train the networks and later test its performance is automatically done by keras, which is the library chosen for this particular project, running with a tensorflow backend. The partition for training was 90% of the data, with the remaining 10% used for testing. 10% of the train set was used for validation.



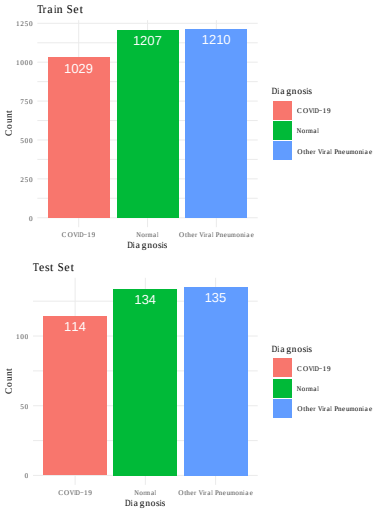


Figure 4: Train and Test partitions of the data set.

The images were subject to **data augmentation** where random alterations within a defined range were performed in order to better generalize to other data sets instead of learning particular features of the images available. Fortunately, radiographies are images taken under protocols defined to have standards for comparison between services and make the images interpretable by healthcare professionals,

so the type of structures and positioning present in the images are fairly stable. In these conditions the alterations performed during augmentation must consider these protocols and respect constraints that come from knowledge in anatomy and clinical evidence. The ones performed were the following:

Rotation: 5 degrees of range, given the positioning of the patient is not free during the exam.

Height and Width distortions: 5% distortions were done in order not to stretch so much the proportions of the thoracic structures.

Zoom: Zoom variations in the range of 10%.

Normalization of values: The values in the matrices of data extracted from the pixels in the image files range from 1 to 255, and for better performance with the algorithm these values were normalized to range between 0 and 1.

No vertical or horizontal flipping of the images were applied, despite being a common practice in image analysis. In the case of radiographies, the anatomical structures are not all symmetrical, and these alterations would not be natural, and would only add an unnecessary source of variability.

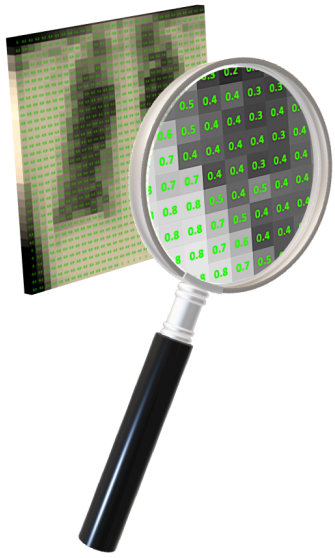


Figure 5: Image arrays are resized and the values are normalized for better training performance.

Methods/Analysis

Two Convolutional Neural Networks were compared. They were chosen because of their dissimilarities. The first model's architecture was humanly defined, while the second was designed with several machine learning techniques autonomously. The first model was trained from scratch, while the second uses the architecture and weights learned previously and transferred to the model shown here for our classification. Finally, the first model is simple and succinct, while the second is quite complex, with intricate details and is the state-of-the-art network for artificial vision, outperforming all the other humanly define models to date.

The images were loaded for training in batches of 32 files, and ran for a maximum of 50 epochs. To prevent overfitting to the train set, training was stopped if after 8 consecutive epochs the validation accuracy didn't improve by at least 1%.

Model 1 - Convolutional Neural Network

The architecture proposed for the first Convolutional Neural Network model consists of 4 convolutional layers as previously described, added a *MaxPool2D* unit layers 2 and 4. Then the arrays are reshaped to a vector (flattened), and fed to a fully connected layer of 256 neurons with ReLU activation, followed by a softmax unit that will provide the prediction for a class.

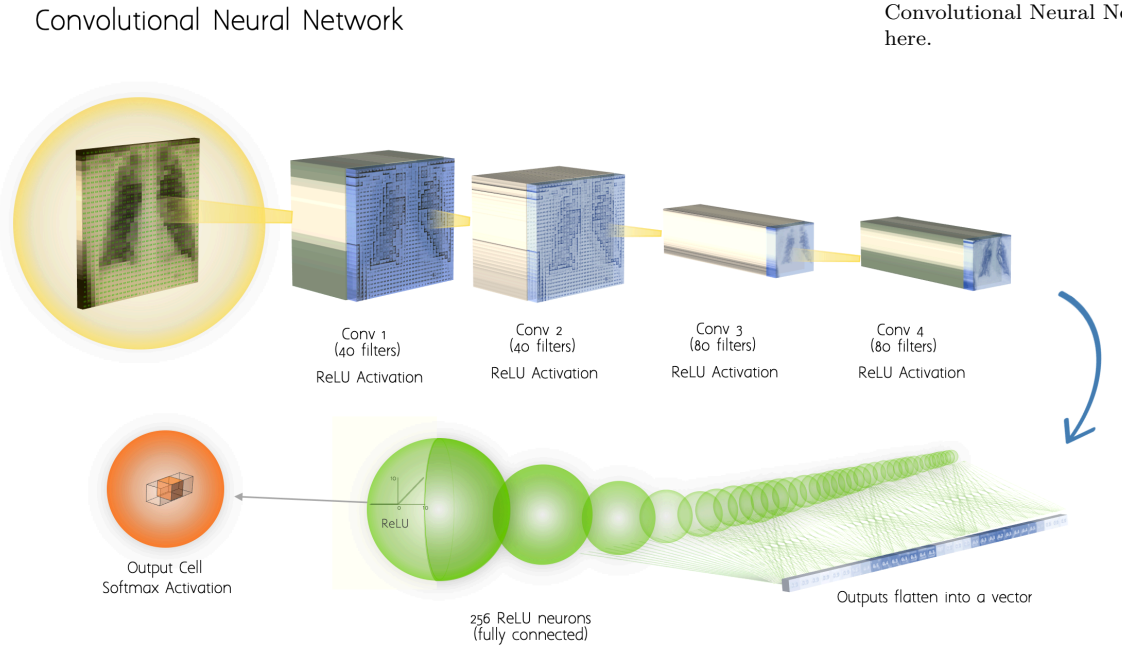


Figure 6: Architecture of Model 1 - Convolutional Neural Network used here.

The model contains 25,190,427 parameters, all of which were trained specifically with the data set presented here. The performance on the test set can be seen in the results section.

Model 2 - Transfer Learning - NASNet

The first model presented was completely defined and designed by the author of this report. What makes this second model interesting is that the architecture was also defined using machine learning techniques, namely reinforcement learning. It was a Neural Architecture Search (NAS)[4] performed via training on more than a million images from the ImageNet database.

The layers in the network consist of Convolutional units, but the method for the choice of the filter sizes, the number of filters, and connectivity used Recurrent Neural Networks to find the best sequence of layers. Their shapes, and aggregations definitions were based on the activation of a Softmax activation. Then the weights for the convolutional layers could be trained using images. The networks that performed better during this architecture definition received the reinforcement for the next iteration of training, which was the accuracy of that architecture on the validation dataset. As a result, the network has learned rich feature representations for a wide range of images. It trained using 800 GPUs simultaneously, to train 12800 models which were evaluated on their validation accuracy after running for 50 epochs each.

What we did here was to take the weights learned by this network and replaced the final layer with new ones to use that massive training to the goal of classifying the dataset. The final architecture of the model has 85,175,125 parameters, of which 258,307 were trained with the data set of X-ray images and the remaining 84,916,818 were transferred from the training on the imagenet data. The image input size must be 331 by 331, and the output was defined to be the 3 diagnosis classes.

The performance on the test set can be seen in the results section.

[4] Based on the neuroscientific literature, the way the connections are tried and reinforced based on the outcomes of the final outputs of the network seems to be the most similar to how the biological neural networks are wired for learning once the basic structural pathways are organized according to code on the DNA of a particular species and for that reason this network was chosen for this project.

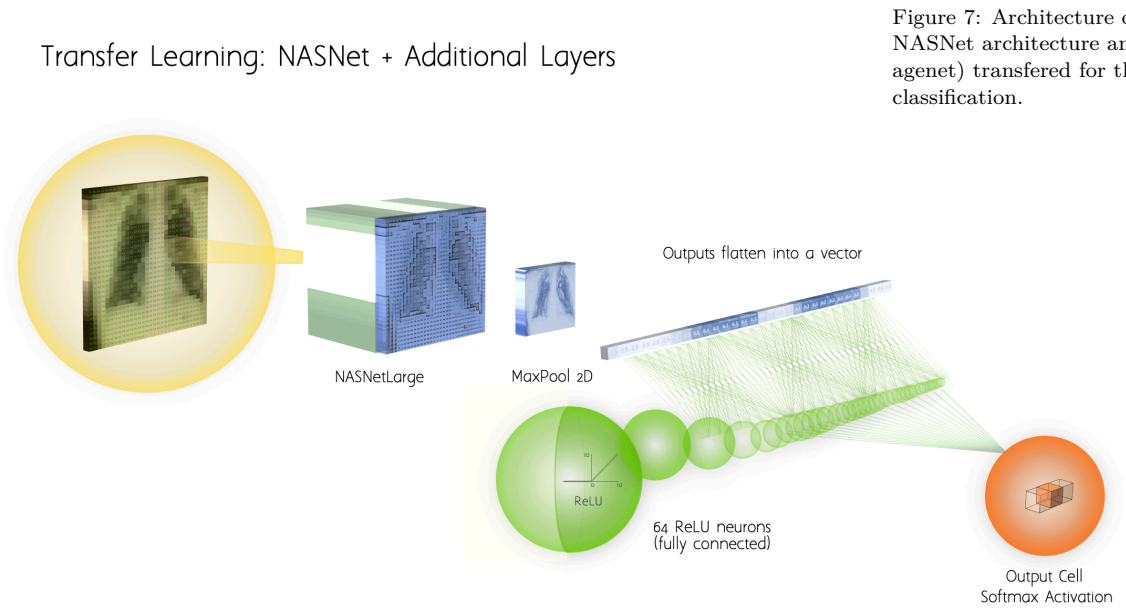


Figure 7: Architecture of Model 2 - NASNet architecture and weights (imagenet) transfered for this diagnosis classification.

Results

The models predicted a $class_i$ for each image. The predictions could be of 4 categories:

TP_{class_i} = True Positive, the correct predictions of the $class_i$.

TN_{class_i} = True Negative, correct predictions of the absence of $class_i$.

FP_{class_i} = False Positive, erroneous predictions of the $class_i$.

FN_{class_i} = False Negative, erroneous predictions of the absence of $class_i$.

Based on the proportions of these outcomes, the following metrics were generated, and were used to compare the models:

$$Accuracy_{class_i} = \frac{TP_{class_i} + TN_{class_i}}{TP_{class_i} + TN_{class_i} + FP_{class_i} + FN_{class_i}}$$

$$Precision_{class_i} = \frac{TP_{class_i}}{TP_{class_i} + FP_{class_i}}$$

$$Recall_{class_i} = \frac{TP_{class_i}}{TP_{class_i} + FN_{class_i}}$$

$$F1\_score = 2\frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

$$AUC_f = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|}$$

where:

$\mathbf{1}[f(t_0) < f(t_1)]$ represents the *indicator function* which returns 1 if the inequality is true and 0 otherwise,

$\mathcal{D}^0$ = set of negative examples (not $class_i$).

$\mathcal{D}^1$ = set of positive examples (belongs to $class_i$).

The table below shows the metrics for both models.

	loss	acc	precision	recall	auc	F1_score
NAS_metrics	0.2352	0.9119	0.9117	0.9091	0.9867	0.9104
CNN_metrics	0.1615	0.9432	0.9432	0.9432	0.9918	0.9432

The simpler Model 1 outperformed Model 2 and that result is counterintuitive at first. Further steps can be taken to improve the performance of Model 2 to reflect the reputation NASNet has in accuracy. Nonetheless, both models show surprisingly good accuracy levels, being above human performance. Also, the sensitivity to detection of COVID-19 using **only the radiography images** was 94.32% for Model 1 and 90.91% for Model 2.

They are comparable to that shown by Vieceli et al.(2020) using 2 laboratory tests and chest radiographies (**96% Recall** and **AUC = 0.827**). The ability to perform the discrimination between the classes measured by the AUC metrics[5] indicates the CNN models also performed better than that study.

[5] The AUC metric was calculated separately for each of the 3 classes, and the results were averaged for a unified score for the model.

Conclusion

The training of CNNs can be rather expensive and time consuming, but the results seem to be worth the investment. The models presented here outperformed the metrics in clinical studies where the classification of COVID-19 diagnosis was done with blood tests that take longer, and need more apparatus for collection and analysis than the radiography. It is possible, after further tests and improvements, that this kind of image analysis could take a more central role in the early screening for the presence of the disease and thus make faster and more accurate diagnosis.

The performance of the simpler CNN was better than the transferred learning from NAS. This may be an indication that the top layers need fine tuning of their parameters, and perhaps alterations in the architecture and those are the next steps recommended for the improvement of the model. It is important to remember that the dataset shown here is split with a similar amount of observations for COVID-19 patients and for all the other viral pneumoniae and normal subjects. That is not the case in a real scenario and the application of such ML techniques to clinical use is far more complicated.