Deep Learning Model to Identify COVID-19 Cases from Chest Radiographs

Matías Cam Arellano, Oscar E. Ramos
Universidad de Ingenieria y Tecnologia - UTEC, Lima, Peru
Email: {matias.cam, oramos}@utec.edu.pe

***Abstract*—The interpretation of radiographs is critical for the detection of many diseases, specially in the thoracic part, which is where *COVID-19* attacks. Many people around the world are suffering from this disease, because of the easy spread of the virus. In an attempt to help physicians in their diagnosis of *COVID-19*, since it can be seen from a frontal view chest radiograph, deep learning approaches have recently been introduced to deal with this detection task. The purpose of this work is to investigate how well current deep learning algorithms perform on the detection of *COVID-19*, and to give hints on how the approach can be used in the future on real clinical settings, to help professional radiologists.**

***Index Terms*—COVID-19, Deep Learning, Medical diagnosis, Convolutional Neural Networks**

I. Introduction

COVID-19 is up to date a global pandemic that have a devastating effect on the health of the population. According to the WHO (World Health Organization), the incubation period of the virus ranges from 2 to 10 days, and the patient can transmit the virus during that period of time, while being completely asymptomatic, and fully unaware of his condition [1]. Early studies found that patients present abnormalities in their chest radiography images, and some special characteristics can be found in those patients who are infected with *COVID-19* [2]. This paper deals with the implementation and design of convolutional neural networks for the detection of *COVID-19* cases from chest X-ray's images obtained from open databases, but can also be implemented in a clinical setting as a tool for doctors to give a more confident diagnose. The hope is that this deep learning solution for detecting *COVID-19* can be use to develop a better algorithm to detect more thoracic diseases and to have a high accuracy rate.

II. Theoretical Framework

A. Class Imbalance Problem

A problem when training neural networks occurs when the class distributions in the training data are highly imbalanced [3], since most classifying algorithms tend to predict the most frequent class and still achieve a high accuracy. This causes infrequent classes to have very low predictions. There exist several ways for dealing with this problem; for instance, using a weighted loss or resampling.

The approach for class imbalance based on a weighted loss consists in multiplying the loss function with a weight that represents the amount of data of the class coming from the dataset. Let the original loss function be a cross-entropy function for two classes denoted as $L(\mathbf{x}, y)$, where $\mathbf{x}$ is the input vector, and y is its corresponding label. The introduction of the weighted-loss approach transforms this cost function to the weighted loss $\hat{L}(\mathbf{x}, y, w_n, w_p)$, which is defined as

$$\hat{L}(\mathbf{x}, y, w_n, w_p) = \begin{cases} -w_p \log p(y=1|\mathbf{x}), & \text{if } y = 1 \\ -w_n \log p(y=0|\mathbf{x}), & \text{if } y = 0 \end{cases},$$

where $p(y=1|\mathbf{x})$ and $p(y=0|\mathbf{x})$ denote the probability of the classification given the input data $\mathbf{x}$, and w_p, w_n represent the imbalance weights associated with the positive and negative classes, respectively. Considering that there are a total of N input samples, N_n negative instances, and N_p positive instances, such that $N_n + N_p = N$, the positive imbalance weight is defined as $w_p = \frac{N_n}{N}$, and the negative imbalance weight as $w_n = \frac{N_p}{N}$.

Another approach to deal with the imbalance problem consists in resampling. This approach clusters the classes and duplicates or eliminates the date until eventually obtaining the same amount of data for each class.

B. Working with Small Datasets

The outbreak of *COVID-19* is fairly recent and, therefore, there are no large datasets to be used in order to properly train a deep neural network from scratch. Due to this constraint, we instead used two alternatives for dealing with small datasets: transfer learning, and data augmentation.

Our first approach consisted in using transfer learning, which is a methodology that uses an already trained convolutional neural network (CNN), but applies a fine tunning to re-train the layers that are closest to the output units. This training of only a few layers is not computationally expensive, does not need a huge amount of data, and is capable of recognizing new complex objects. The layers that are kept intact are used as feature detectors for the input images, in this case. For the radiography images our approach will be based on the pre-trained DenseNet121 [4] network, whose architecture is shown in Fig. 1.

The other approach that we undertook to deal with the small amount of data is *data augmentation*. This technique consists in generating several instances from a single instance, by performing some small modifications of it. In the case of images, which is the case that we deal with in this work,

978-1-7281-9377-9/20/$31.00 ©2020 IEEE

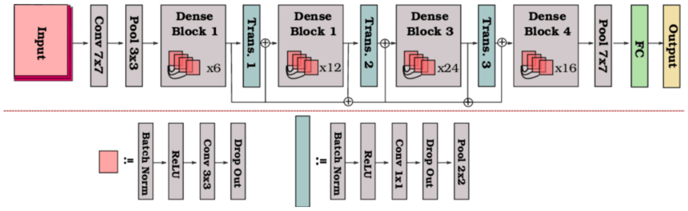

Fig. 1. DenseNet121 Architecture

the small modifications are non-severe affine transformations and brightness modifications. For instance, there can be small rotations such as at most 30° to the right or to the left, and small brightness changes as long as they do not fully obscure the image. Other transformations, such as mirroring the image, have to be carefully considered, since when analyzing chest X-rays it might happen that the sickness only occurs in the left side or right side of the chest.

C. Evaluation Metrics

Let the real label of an instance be y, and let its prediction be $\hat{y}$, where the values they can take on are 1 for a patient with the disease, or 0 for a patient without the disease. To evaluate the prediction model in classification, it is important to define the terms that are referred to as true positives $TP = p(\hat{y}=1|y=1)$, false positives $FP = p(\hat{y}=1|y=0)$, true negatives $TN = p(\hat{y}=0|y=0)$, and false negatives $FN = p(\hat{y}=0|y=1)$. The true values (TP and TN) correspond to the instances, positive or negative, that have been correctly classified. The false values (FP and FN) correspond to those instances that have been wrongly classified; for instance, FP means that the instance was classified as positive but its actual label is negative. The evaluation metrics are based on the computation of these values.

The *accuracy* of the classification is defined as the probability of a correct prediction and can be expressed as

$$p(\hat{y}=1|y=1)p(y=1) + p(\hat{y}=0|y=0)p(y=0)$$

where, $p(\hat{y}=1|y=1)$ is also known as the *sensitivity*, $p(\hat{y}=0|y=0)$ is also known as *specificity*, and $p(y=1)$ is also known as the *prevalence*. These probabilities can be obtained from the ratio between the true and false positives and negatives. Using this approach, the accuracy can be computed in practice as $\frac{TP+TN}{N}$, where $N = TP + TN + FP + FN$ is the total number of instances. The *prevalence*, which is also important in the medical field since it can be interpreted as the ratio of people in the population who already have the disease, can be computed as $\frac{1}{N}\sum_i y_i$, where $y_i = 1$ when the example is positive or has the disease, and 0 otherwise. On the other hand, the sensitivity and specificity, also important for diagnostic tests, are computed as $\frac{TP}{TP+FN}$, and as $\frac{TN}{TN+FP}$, respectively.

For the medical field, in order to determine whether the patient has the disease or not, the so-called *positive and negative predicted values (PPV and NPV)* are used. These metrics are computed as $PPV = \frac{TP}{TP+FP}$ and $NPV = \frac{TN}{TN+FN}$. On the one hand, PPV is defined as the probability that following a positive test result, the individual will truly have that specific disease. On the other hand, NPV is defined as the probability that following a negative test result, that individual will truly not have that specific disease [5].

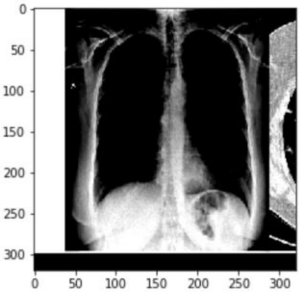

Fig. 2. Frontal View of chest radiography

D. ROC curve (Medical)

The measurement of the area under the ROC curve, also known as AUCROC or C-statistic, provides a way to evaluate the goodness of a fit. In medical terms, this number provides the probability that a patient who experienced a condition randomly selected had a higher risk score than a patient who is without any experiences of the condition. The ROC curve is used to summarize the model output across all thresholds, and provides a good sense of the prediction of a given model.

III. Methodology

A. Data for Training

The databases that were used to train the deep learning algorithm are the *COVID-19 Radiography* database [6], as well as the *ChestX-ray14* dataset [7]. Both datasets are currently the largest public repositories on *COVID-19* and chest radiographs, containing over 100000 frontal view images of lung diseases. The *ChestX-ray14* dataset contains several types of diseases, and this work uses a DenseNet121 pre-trained model on this dataset, which is not targeted at detecting *COVID-19* but other lung diseases. As mentioned before, our approach consists in applying transfer learning to retrain the last layer of the convolutional neural network in order to predict possible *COVID-19* positive or negative cases. An example of an image used to train, validate and test the model is shown in Fig. 2.

B. Class Imbalance for COVID-19

One of the main challenge when working with the medical diagnosis of *COVID-19* is the large class imbalance that the image dataset presents. Since *COVID-19* is a new disease, datasets are not rich in data of positive instances, and the access to local data (databases from local private or public hospitals) is currently very restrictive. In our case, the frequency of each class in that database that is being used is shown in Fig. 3 (left). From this figure, it can be clearly seen that the negative cases are significantly more frequent than the positive cases. More concretely, the negative cases (healty

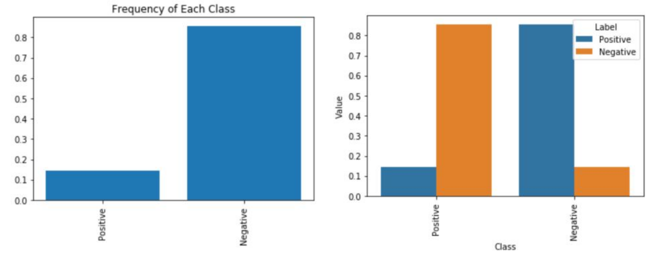

Fig. 3. Frequency of each class in the dataset (left), and contributions of each class in the same dataset (right)

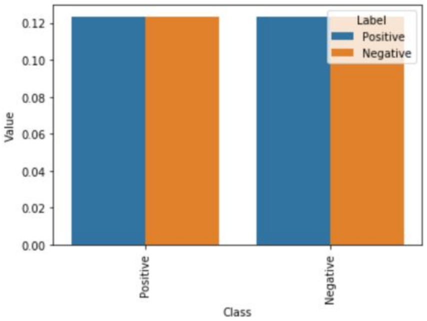

Fig. 4. Contributions balanced

patients) constitute around 85% of the dataset, whereas the positive cases that need to be identified only constitute the 15% of all the cases.

The positive and negative frequencies, which will be referred to as f_p and f_n respectively, can be computed as shown in Fig. 3 (right), based on the previous information. Using these frequencies, the model can extract the imbalance weight, considering that $w_p f_p = w_n f_n$. Then, these weights can be finally obtained as $w_p = f_n$, and $w_n = f_p$. Using these weights, the model loss function can be balanced for the contribution of positive and negative labels. This balancing process can be verified computing the new frequencies, as Fig. 4 shows. From this figure, it can be seen that by applying proper weightings to the positive and negative labels in the class, the effect of each one on the model would represent the same aggregate contribution to the loss function.

C. Training

As mentioned before, this work uses a weighted loss function to return a final loss function that calculates the weighted losses for each training batch. We also add a small value, called ϵ, to the predicted values before taking their $\log$. The rationale for adding this value is to avoid numerical errors that can occur if the predicted value is zero.

Our approach consists in using a pretrained DenseNet121 model [4], which has been trained using the aforementioned datasets. For addressing the specific situation of the detection we are interested in, we add two layers on top of the base network. These two layers are the following.

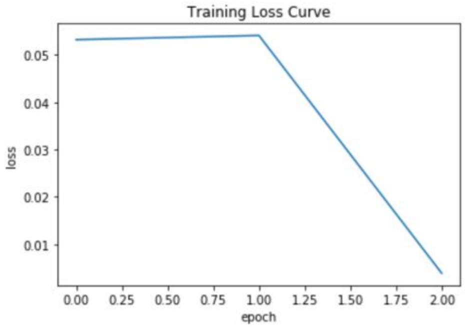

Fig. 5. Training Loss Curve

- The first layer is a global average pooling layer, that will be used to get the average of the latest convolution layer directly coming from the DenseNet121 model.
- The second layer is a fully connected dense layer with a sigmoid function used as activation, since we are dealing with a two-class problem.

In our case, we train this last layer, and use the rest of the pre-trained convolutional neural network as a feature detector. Since it has been trained on different diseases, but all of them coming from chest radiographs, it is expected that the network will provide good and distinctive features that will be useful for our later training.

IV. Results

The trained and tested the system on a desktop computer using Python and some packages available for data science and deep learning such as numpy, pandas, seaborn, matplotlib, Keras, among others. Due to the computational limitations, we performed mini-batch gradient descent using a few epochs. The training loss that was obtained with this approach is shown in Fig. 5. It can be seen that the goal of the training is being achieved, since the loss function is decreasing. Note that the curve seems to be decreasing and suddenly stop, but this is an effect of how the graph is being shown: only the loss at the very end of each epoch is being plotted. A deeper analysis on the results within each epoch shows that the final minimum has been, in fact, achieved before the last epoch and the shown value at the end of the epoch is just a single sample. An example of a classified image is shown in Fig. 6: the right image shows a patient with *COVID-19*, and the left image a patient without this disease. From these images, it can be seen that it requires radiological training to distinguish both case, but the developed system is also able to perform this task.

The first evaluation of the results of the proposed classification method on the test set is the confusion matrix shown in Table I. As depicted, the amount of false positives and false negatives is small compared to the true positives and true negatives. This can be verified provided that the accuracy of the model is 94.7%. Also, using the information of the matrix, it can be found that the prevalence value is 0.5, which states that half the dataset used for the test has the disease

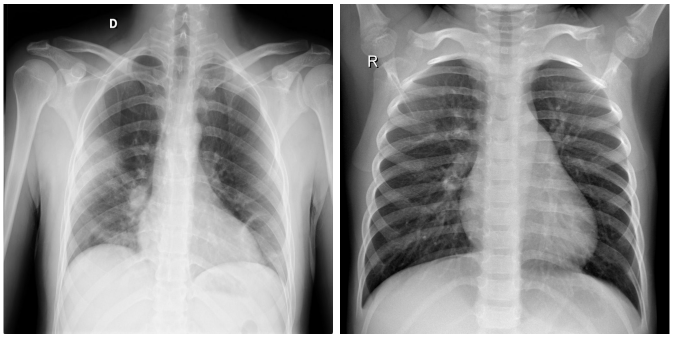

Fig. 6. Radiography images of a patient with COVID-19 (left) and without it (right)

and the other half does not have the disease or is a patient that presents a normal condition.

TABLE I
CONFUSION MATRIX

Label	Positive prediction	Negative prediction
Patient with COVID-19	17	2
Patient without COVID-19	0	19

The sensitivity and specificity of the model are shown in Table II both for the patients with the disease and for those without the disease. Note that these results do not depend on the prevalence of the diseased class in the dataset, because the statistics are computed within the same class. That is, the sensitivity only considers the outputs of the diseased class while the specificity only considers the outputs of the normal class.

TABLE II
SENSITIVITY AND SPECIFICITY TABLE

Label	Sensitivity	Specificity
Patient with COVID-19	0.894737	1
Patient without COVID-19	1	0.894737

Since this work is focused on a medical application, it is useful to evaluate the PPV and NPV values. The results for each case are shown in table III. It can be seen that predictions are at around 90% for each class, having low false prediction cases, which is a promising medical indicator for the proposed model.

TABLE III
PPV AND NPV TABLE

Label	PPV	NPV
Patient with COVID-19	1	0.904762
Patient without COVID-19	0.904762	1

The ROC curve of the proposed model for classification is shown in Fig. 7. As can be seen, the obtained curve is close to the ideal ROC curve, with 0.947 for positive tested cases, and 0.96 for negative tested cases. This results shows that the trained model works across different threshold values, and that the prediction outputs are mostly correct.

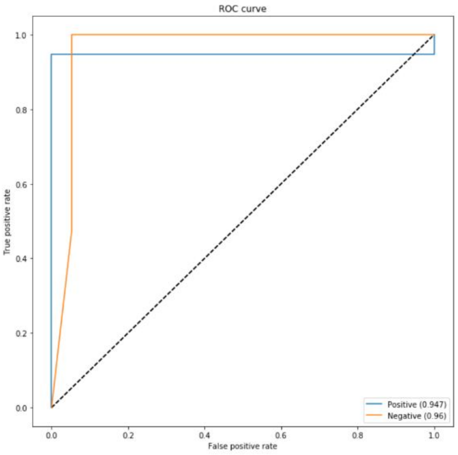

Fig. 7. ROC curve

V. CONCLUSION

The proposed methodology for using deep learning in this paper has been shown to have a very high performance that can detect *COVID-19* virus in frontal view chest radiographs. This technology can be used as a tool to improve and to give the security to the physicians while diagnosing a disease, and moreover under the current circumstances.

However, further studies are going to be required to make the approach work with other thoracic pathologies, and to determine the feasibility of these results under a controlled clinical setting. *COVID-19* is a fairly new disease, so more data is also required to keep on working on the model an developing a better model to be used nowadays.

REFERENCES

[1] W. H. Organization. (2020) Novel coronavirus (2019-ncov). [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200127-sitrep-7-2019–ncov.pdf

[2] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, “Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review,” *Clinical Imaging*, 2020.

[3] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018.

[4] N. Radwan, “Leveraging sparse and dense features for reliable state estimation in urban environments,” Ph.D. dissertation, University of Freiburg, Freiburg im Breisgau, Germany, 2019.

[5] H. B. Wong and G. H. Lim, “Measures of diagnostic accuracy: sensitivity, specificity, ppv and npv,” *Proceedings of Singapore healthcare*, vol. 20, no. 4, pp. 316–318, 2011.

[6] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi *et al.*, “Can ai help in screening viral and covid-19 pneumonia?” *arXiv preprint arXiv:2003.13145*, 2020.

[7] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz *et al.*, “Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists,” *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.