INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Arthur Fernandes de Morais

HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS

Final Paper
2021

Course of Computer Engineering

CDU 681.3:16

Arthur Fernandes de Morais

HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS

Advisor

Prof. Dr. Marcus Henrique Victor Júnior (ITA)

Co-advisor

Prof. Dr. Marcos Ricardo Omena de Albuquerque Maximo (ITA)

Co-advisor

Prof[a]. Dr[a]. Mônica Mitiko Soares Matsumoto (ITA)

COMPUTER ENGINEERING

São José dos Campos
Instituto Tecnológico de Aeronáutica

2021

Cataloging-in Publication Data
Documentation and Information Division

Fernandes de Morais, Arthur
HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS / Arthur Fernandes de Morais.
São José dos Campos, 2021.
46f.

Final paper (Undergraduation study) – Course of Computer Engineering– Instituto Tecnológico de Aeronáutica, 2021. Advisor: Prof. Dr. Marcus Henrique Victor Júnior. Co-advisor: Prof. Dr. Marcos Ricardo Omena de Albuquerque Maximo.

1. Otimização. 2. Navegação autônoma. 3. Imagem Termal. 4. Odometrica visual. 5. Algoritmo Genético. 6. Detector de feature. 7. Detector de pontos característicos. I. Instituto Tecnológico de Aeronáutica. II. Title.

BIBLIOGRAPHIC REFERENCE

FERNANDES DE MORAIS, Arthur. **HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS**. 2021. 46f. Final paper (Undergraduation study) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSION OF RIGHTS

AUTHOR'S NAME: Arthur Fernandes de Morais
PUBLICATION TITLE: HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS.
PUBLICATION KIND/YEAR: Final paper (Undergraduation study) / 2021

Arthur Fernandes de Morais
Pedro de Toledo, 205, ap102
12.243-740 – SP/ Brasil

HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS

This publication was accepted like Final Work of Undergraduation Study

Arthur Morais

Arthur Fernandes de Morais

Author

Marcus Henrique Victor Júnior (ITA)

Advisor

Marco Maximo

Marcos Ricardo Omena de Albuquerque Maximo (ITA)

Co-advisor

Monica M.

Mônica Mitiko Soares Matsumoto (ITA)

Co-advisor

Marco Maximo

Prof. Dr. Marcos Ricardo Omena de Albuquerque Maximo
Course Coordinator of Computer Engineering

São José dos Campos: NOVEMBER 12, 2021.

I dedicate this work to my father, Arimateia, and my mother, Nara, because they made me believe that education is the most powerful tool for human freedom.

Acknowledgments

This work means more than the words and ideas exposed here: it is the end of one of the most important cycles of my life.

First of all, I would like to thank my family, who always believed in me, especially in difficult moments, to friends who became my second family and that I would never have made it this far without them.

To all my teachers, especially Tibério Beserra, my first computer teacher, that was never content to teach just what was just necessary things.

"Have the courage to follow your heart and intuition
They somehow already know what you truly want to become
Everything else is secondary"
— Steven Paul Jobs

Resumo

Durante as últimas décadas nos apoiamos sobre a hipótese de que nosso método de raciocínio lógico é justificadamente acurado para desenvolvermos algoritmos computacionais análogos ao funcionamento de nosso intelecto. Assim, com essa mesma lógica, é natural surgir a dúvida se esses mesmos algoritmos estão de fato tomando decisões coerentes com o ambiente ao qual foram treinados. A necessidade de obter justificativas plausíveis com uma dada decisão é tão maior quanto o risco associado à essa decisão. Aliado a isso, temos visto cada vez mais aplicações computadorizadas inseridas em ambientes médicos devido os benefícios atrelados a se utilizar inteligência artificial para reduzir erros clínicos. Muito embora nós tenhamos ampliado a expectativa de vida humana ao redor do mundo, ainda estamos sujeitos a pandemias globais que afetam principalmente os menos favorecidos, como ficou claro durante a pandemia do COVID-19 iniciada em 2019. Inspirado em todo esse contexto, esse trabalho se propõe a explorar uma técnica de explicabilidade que não depende da arquitetura da rede neural artificial utilizada em um treinamento de aprendizado de máquina. Mais especificamente, por meio da técnica do Grad-CAM, foi possível analisar a tomada de decisão de uma rede neural para o diagnóstico de SARS-COV-2. A rede neural desenvolvida obteve 92.8% de acurácia no diagnóstico de COVID-19 a partir de 13.993 tomografias computadorizadas de pulmões de pacientes de diferentes países. Apesar do escopo do trabalho, convém salientar que a técnica utilizada pode ser aplicada em outros campos além do domínio da medicina.

Abstract

During the last few decades, we have supported the hypothesis that our method of logical reasoning is justifiably accurate in developing computational algorithms analogous to the workings of our intellect. Thus, it is natural to doubt whether these same algorithms are making decisions consistent with the environment with this same logic. The need to obtain plausible justifications for a given decision is as more significant as the risk associated with that decision. Allied to this, we have seen more and more computer applications inserted in medical environments due to the benefits of using artificial intelligence to reduce clinical errors. Even though we have increased human life expectancy worldwide, we are still subject to global pandemics that primarily affect the poor, as became apparent during the COVID-19 pandemic that started in 2019. Inspired by this entire context, this work proposes to explore an explicability technique that does not depend on the artificial neural network architecture used in machine learning training. More specifically, through the Grad-CAM technique, it was possible to analyze the decision-making of a neural network for the diagnosis of SARS-COV-2. The developed neural network obtained 92.8% accuracy in diagnosing COVID-19 from 13,993 computed tomography scans of lungs from patients from different countries. Despite the scope of the work, it should be noted that the technique used can be applied in other fields beyond medicine.

List of Figures

List of Abbreviations and Acronyms

CNN	Convolutional neural network
CT	Computer tomography
COVID-19	Coronavirus 2019 disease
ML	Machine Learning
SARS-Cov-2	Severe acute respiratory syndrome

Contents

1 Introduction

In this section, we verify the recent advances in artificial intelligence techniques and their impact on the daily life of contemporary society. Furthermore, even with enormous benefits, the use of these techniques today has not been enough to prevent global health pandemics.

1.1 Motivation

1.1.1 Artificial intelligence

In the first lines of the article "Computing Machinery and Intelligence," Alan Turing in 1950 proposed the question: Can machines think?

The evolution of computer science from that question to the point where recommendation algorithms are so customized and optimized that it would surprise any science fiction writer of the early twentieth century is evident.

Not only in computer science but for thousands of years, how human beings learn remains an open question. Despite the most solid theories, possibly we will never fully know how the cognitive process of human beings (GARDNER, 2004).

However, even with the limitations of what we know about our learning process, the beneficial impact of computational solutions that mimic what we understand as intelligence is evident. Autonomous agents have gone from autonomous chess players to rockets capable of landing on the terrestrial ground after re-entry into the atmosphere.

It is possible to highlight some recent social phenomena that allowed us to advance in the construction of intelligent systems. Firstly, hardware development enabled a growing evolution in computational power, including computers considered for personal use, as shown in Figure 1.3.

Furthermore, in a society increasingly connected to mobile devices that generate data, broad access has made experimentation and information exchange easier in the scientific community. Consequently, it was possible to develop increasingly accurate algorithms.

FIGURE 1.1 – The Supercomputer Deep Blue played against world chess champion Garry Kasparov in 1996. The supercomputer was the first to beat a world chess champion in a tournament with official time rules.

FIGURE 1.2 – The Falcon 9 rocket designed by Space X Company is considered the first capable of re-entering the atmosphere and landing vertically autonomously. The first successful re-entry was in December 2015.

Although there are similarities between different possible approaches to building machine learning algorithms, it is necessary to define the problem to be solved well to select the best possible strategy. In this sense, in search of developing intuitive ways to identify patterns in images, the scientific community has developed a field called computer vision.

Historically researchers were developing Computer vision algorithms from biological inspiration, which is nothing new in scientific discoveries. Just as the prototypes of airplanes have experimented with structural analogies of animals capable of flying, systems capable of identifying patterns in images were developed from the fundamental morphological analysis of our brain, particularly our visual cortex.

However, following the same analogy, we have historically found that the planes most likely to solve the problem of exploring airspace as a means of transport do not need to flap their wings. As well as the fundamental structures of robust computer vision algorithms do not necessarily need to mimic our neurons. Even because the field of neuroscience is relatively new, and exploring the morphological functioning of our brains comes up against ethical issues (FUCHS, 2006).

In this sense, in 2006, Geoffrey Hinton published an article known as state of the art about pattern recognition (HINTON *et al.*, 2006). Hinton and his solution to the problem of handwritten digit recognition, with more than 98 percent accuracy, inaugurated the emergence of a field of research called Deep Learning.

Recent advances in deep learning techniques (MINAR *et al.*,) associated with the development of increasingly efficient hardware have democratized access to the development of more innovative applications. In this context, Convolutional Neural Networks stand out, a technique considered to be the current state of the art for image recognition applications.

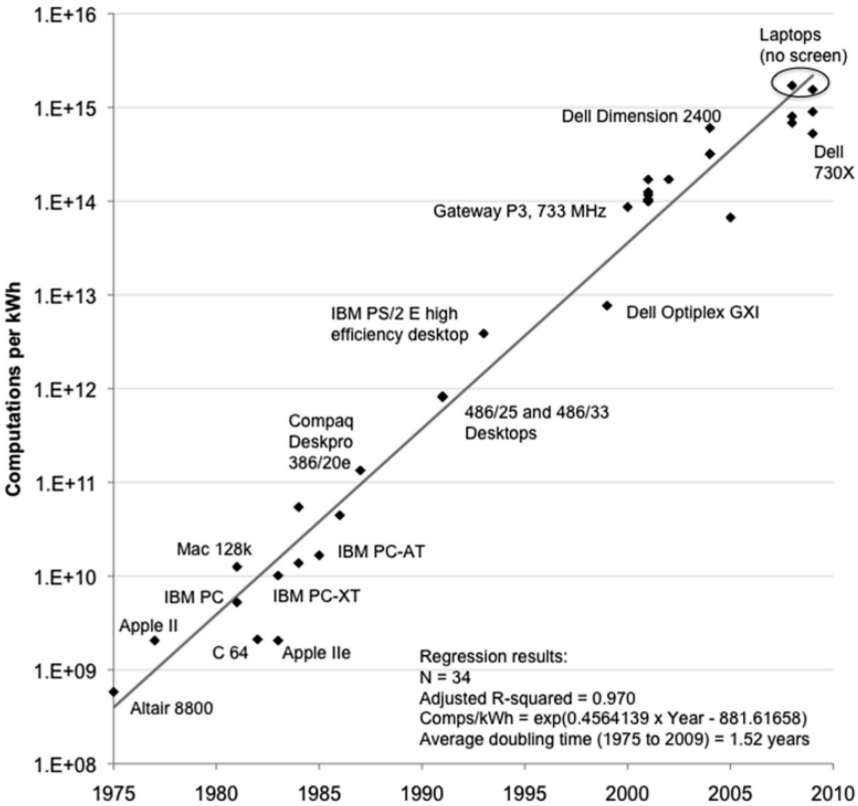

FIGURE 1.3 – Time series of Computations per kilowatt-hour for personal computers. Efficiency doubled every 1.52 years from 1975 to 2009 (KOOMEY *et al.*, 2011).This fact was essential for classical machine learning algorithms development because several iterations are necessary for satisfactory learning in this context.

However, it is natural for a pessimistic idea to emerge in this optimistic context. We incompletely know what makes it possible for us to learn, and we undeniably make decisions that can logically be considered irrational. Suppose we replicate a part of the cognitive process that we think we understand to computational algorithms, hoping that they make intelligent decisions. In that case, it is natural to doubt automatic decision-making. That is worst in environments where the human decision is questionable.

We are in this environment of uncertainty. On the one hand, in practice, we are increasingly faced with intelligent systems capable of facilitating our daily lives and freeing us from repetitive tasks. On the other hand, this same advance in fields that involve ethical issues, such as medicine, cannot be adhered to without a more significant effort to validate the feedback behavior of these decisions over time.

Then even if Convolutional Neural Networks (CNN) are in state of the art for image recognition, the interpretability of the results obtained using CNNs for image classification is difficult. Consequently, there is a risk that applications with high accuracy in the tests performed are not learning what they should be. That is, their generalizability becomes undermined.

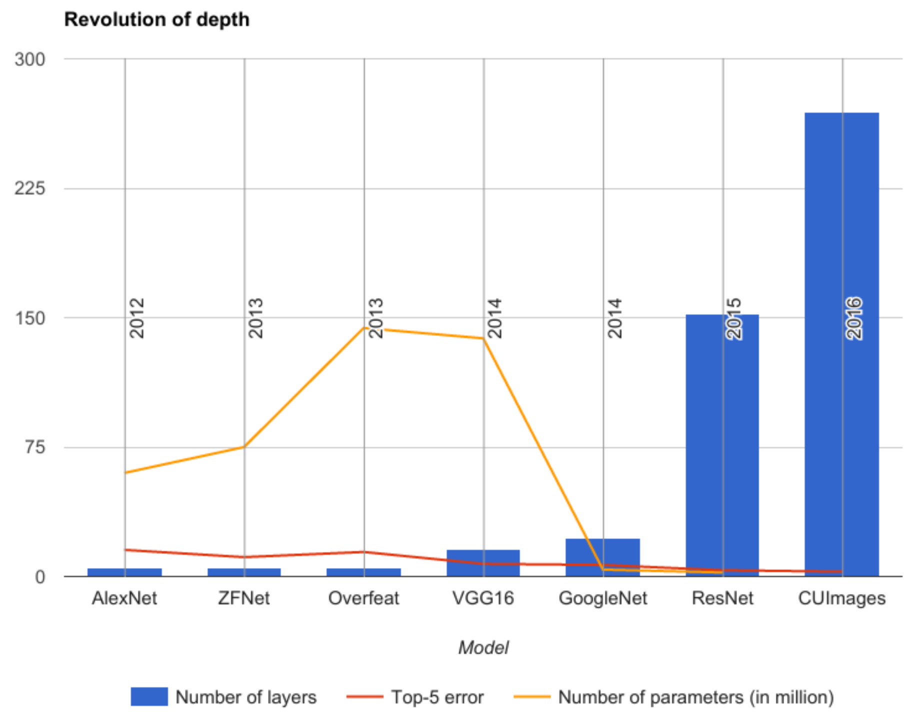

FIGURE 1.4 – An example of the performance improvement of state-of-the-art algorithms in classification tasks is the on-human performance launched by Convolutional Neural Networks (CNNs) in the ImageNet challenge (RUSSAKOVSKY *et al.*, 2015).

To overcome this type of problem, several techniques for explaining the decision of algorithms were developed, including GRAD-CAM. With this type of approach, it is possible to validate the generalization of an already developed model and help during the development of image classifiers that use CNNs.

Given this addendum, it is undeniable that one of the most motivating facts for using and disseminating algorithms that use Machine Learning (ML) is the high speed of evaluation. This feature is especially welcome in critical public health situations, as the delay and inefficiency of health systems can lead to the death of people. Unfortunately, our society has had to live again with a global public health crisis in recent years, which has forced us to review specific concepts about our ability to respond to the environment according to the medical tools available so far.

1.1.2 COVID-19 pandemic

Historically, during the 20th century, the great world wars are commonly highlighted as the most significant risk to the health of humanity at the time. However, more damaging than the possibility of major armed military events in the current century may be the highly contagious diseases caused by microorganisms. Unlike declared armed events, the spread of diseases and a possible pandemic are still outside the political-strategic planning of many countries worldwide.

Because of the significant evolution of access to transport, emphasizing air transport, what in the past could only be a contagious disease easily controllable in a reduced geographic space, can now become a pandemic with a high potential for mutation over time. In this context, the recent pandemic caused by the COVID-19 virus stands out.

Coronavirus 2019 disease (COVID-19) that triggers Severe Acute Respiratory Syndrome (SARS-Cov-2) has already has already caused the death of more than 5 million people, as we can see in Figure 1.5, many of whom have never had a previously reported chronic disease history.

In addition, global financial losses could reach about 8.5 trillion dollars, and approximately 34 million people have fallen into extreme poverty (UN, 2020).

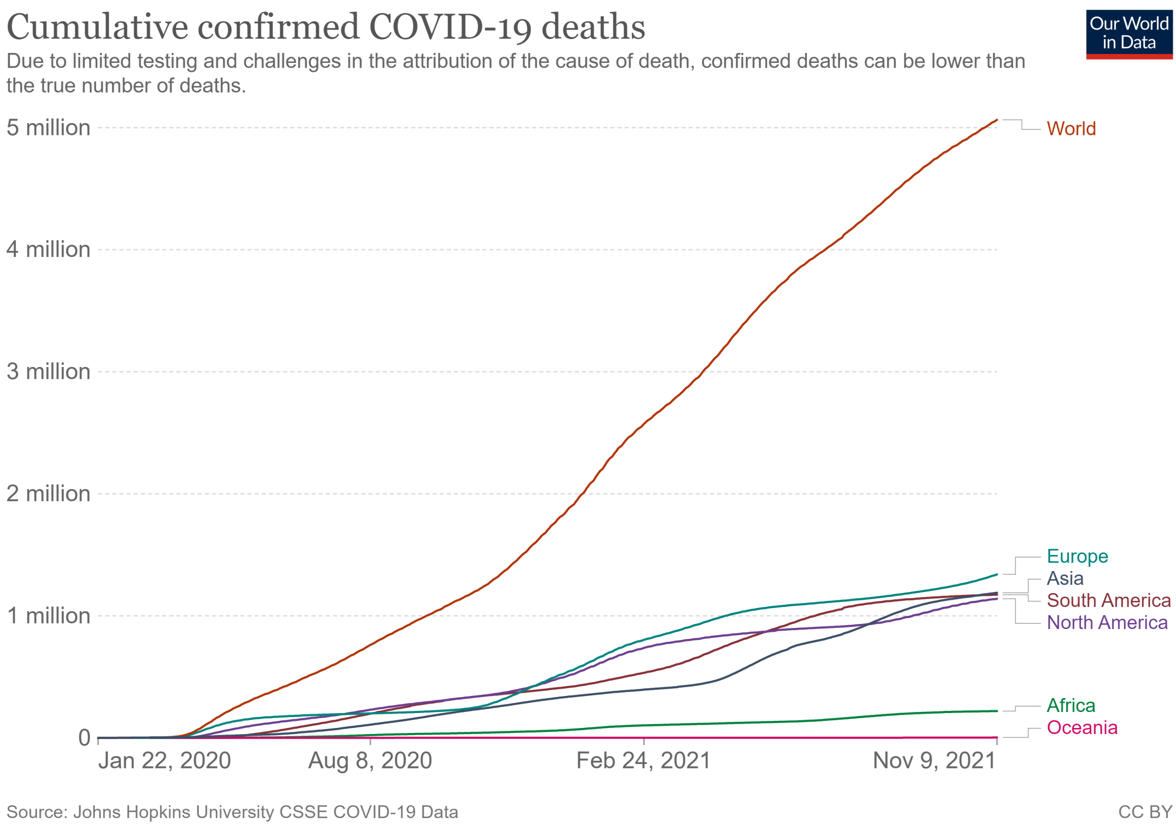

FIGURE 1.5 – Time series of the number of deaths resulting from COVID-19 accumulated for each continent. And the total number of deaths in the world (DATA, 2021).

The need to develop increasingly efficient health systems and adapt to sudden changes in context is obvious. Several deaths by COVID-19 was caused by the discrepancy of social and economic conditions of countries as can be seen by the contrast in the total number of deaths per million people even in countries on the same continent in Figure 1.6.

Among the commonly used methods for diagnosing this disease stand out Reverse

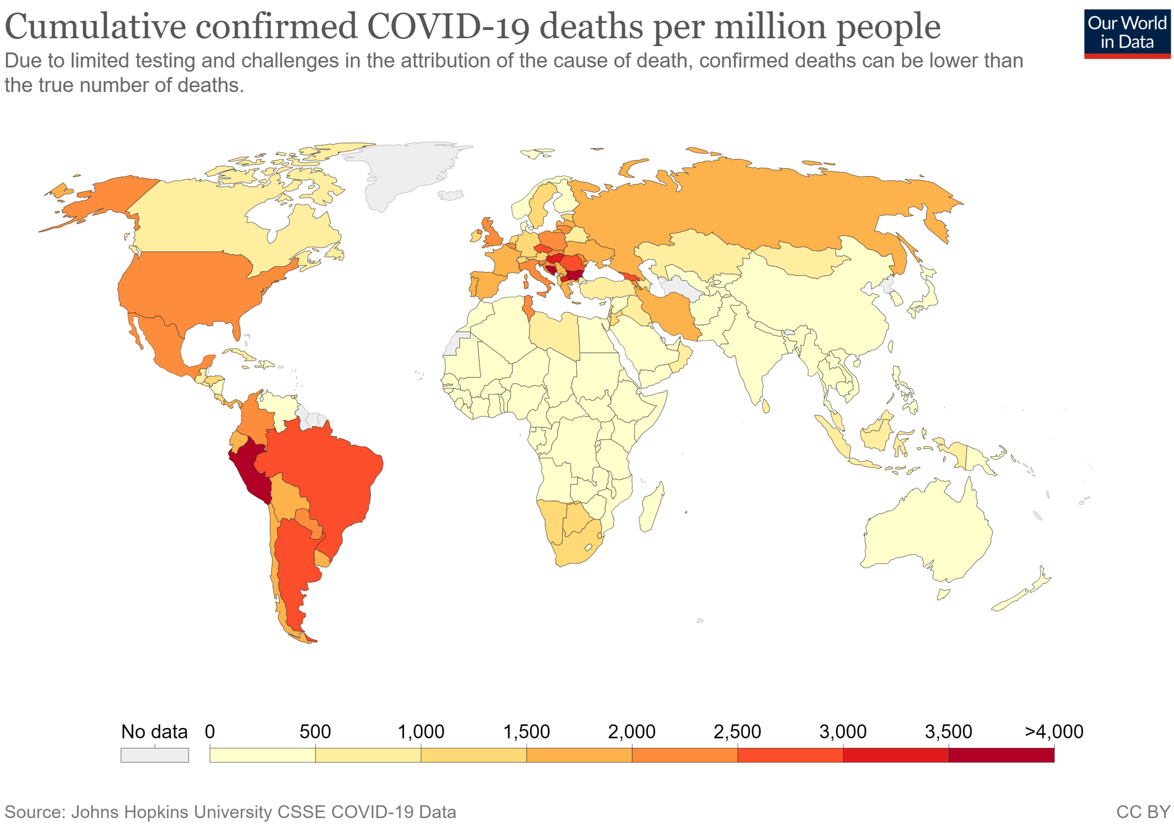

FIGURE 1.6 – Deaths per million people in each country until November 2021. Contrast metrics for neighboring countries indicate infrastructure discrepancies between nations (DATA, 2021).

Transcription Polymerase Chain Reaction (RT-PCR), Computed Tomography analysis Scan (CT-Scan), and Chest X-Ray (CXR) analysis.

Among these, the one that takes the most time for diagnostic is RT-PCR. Thus, the alternative of imaging diagnosis for this disease is a possible solution since it is a non-invasive technique and allows monitoring the evolution of the patient's clinical condition as seen in Figure 1.7.

1.2 Objective

The present work intends to demonstrate the use of a post-hoc explainability technique of Convolutional Neural Networks. More specifically, we applied this technique to a dataset from CT scans of people suspected of being diagnosed with COVID-19. We intend to demonstrate that this technique can assist in developing machine learning algorithms. Additionally, the heat maps generated by this technique possible can help better medical decision-making, reducing the human error rate in this type of diagnosis.

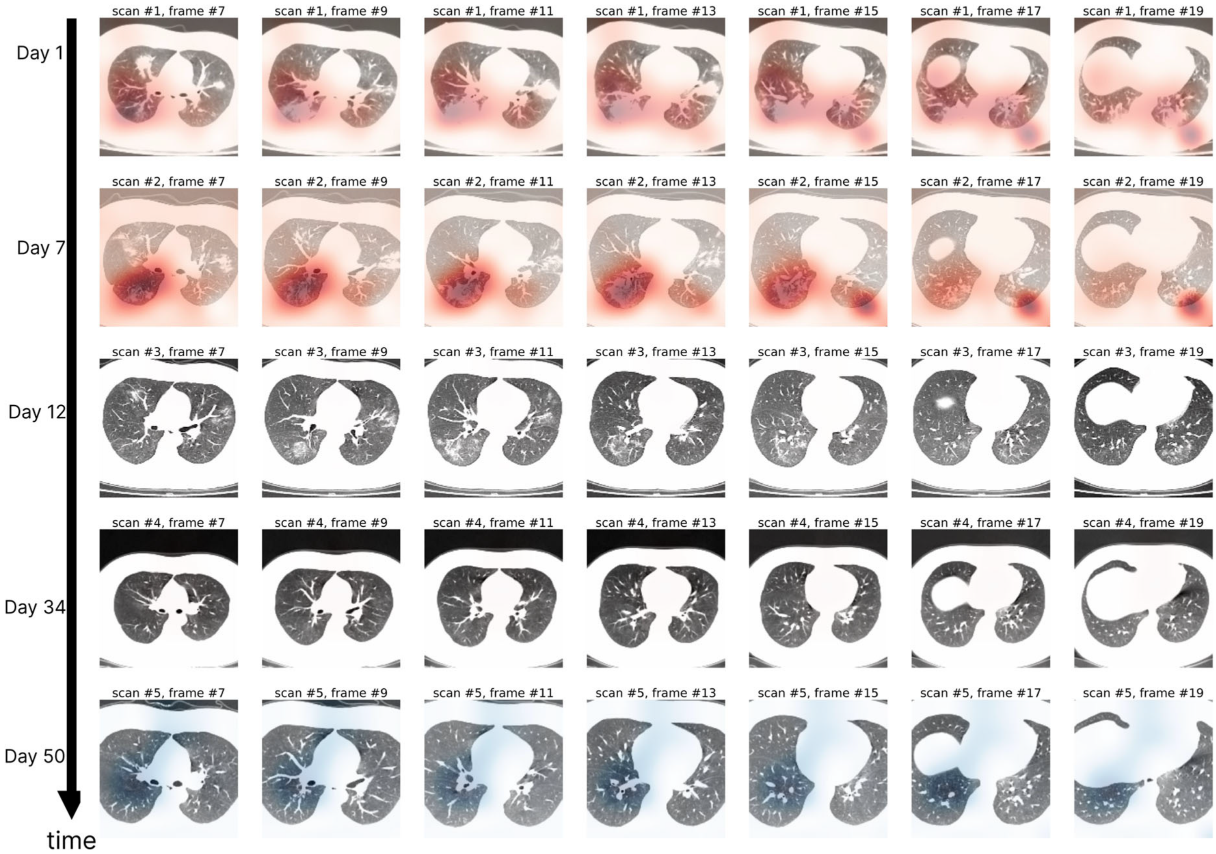

FIGURE 1.7 – Non-invasive follow-up of the evolution of the treatment of patients with COVID-19. The Grad-CAM technique can serve to indicate the improvement in the health status of patients as an alternative diagnosis (SELVARAJU *et al.*, 2016).

1.3 Organization of this work

This work has the following structure:

- Chapter 2 will explain the current related works that served as the basis for developing our research.
- Chapter 3 will describe the origins of ideas developed in machine learning in the computer field of vision and the theory around Deep Learning.
- Chapter 4 will describe selected methods used in experimentation.
- Chapter 5 will describe the results.
- Chapter 6 will describe the conclusion and a brief future vision of the work.

2 Literature review

In this chapter, we intend to highlight the prominent publications related to our study. The research field of Explainable Artificial Intelligence (XAI) is relatively new compared to classical techniques developed for machine learning, but scientific research in the area is growing fast.

2.1 Explainable techniques

This section highlights generalist works aimed at the explanatory techniques of models without necessarily applying them to the diagnosis of COVID-19.

2.1.1 Definitions and concepts of explainability

Even with the growing concern to understand decision-making in intelligent systems, we often cannot identify a clear boundary between each term used in the scientific community. In Vilone e Longo (2020), we find an extensive study that proposes categorizing in a hierarchical way the techniques of Explainable AI developed so far.

One of these techniques is the Grad-CAM technique (SELVARAJU *et al.*, 2016). Our study explicitly uses this approach, as it is helpful for classification tasks. With Grad-CAM, it is possible, in the same input, to generate attention maps for each of the defined classes. Furthermore, a great advantage of this method is its post-hoc characteristic. In other words, it does not change the properties of previously trained neural networks.

Even though feature maps can intuitively help decision-making in many tasks, Adebayo *et al.* (2018) demonstrates that the most popular models so far are not fully developed. As feature maps are generated from previously trained neural networks, depending on how noisy the inputs of these networks are, the map output loses the expected meaning.

Based on studies that demonstrate the inability of these post-hoc care models, (GHASSEMI *et al.*, 2021) presents arguments for the dangers arising from the use of these techniques by non-specialists. This work defends the thesis that, due to the intrinsic limitations of these techniques, it would be more helpful to focus on their applicability during the development of Deep Learning models and not to help the diagnosis from these systems.

2.2 Explainability applied to the diagnosis of COVID-19

This section highlights works that used XAI applied in the specific context of the imaging diagnosis of SARS-COV-2, either in X-ray images or in CT-Scan images.

2.2.1 Grad-CAM applied to COVID-19 diagnosis

Wang *et al.* (2020) is another study that compares the decision-making of a Deep Learning model aided by the Grad-CAM technique to X-Ray images. The study compared the performance of the model and the heat maps generated with diagnoses carried out by specialists.

Lee *et al.* (2021) demonstrates that the Grad-CAM technique can validate COVID-19 classifiers based on computed tomography images of patient lungs. Furthermore, as seen in 2.1, the Grad-CAM technique can be helpful even for 3D analysis.

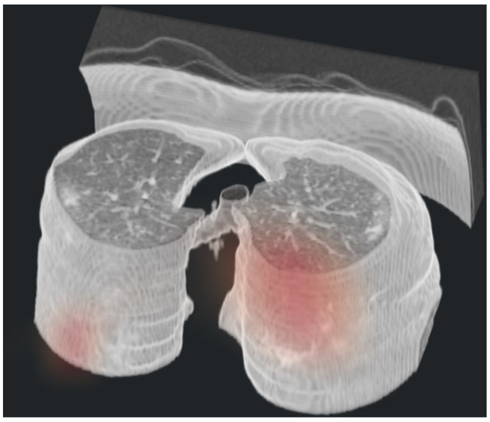

FIGURE 2.1 – 3D Grad-CAM view superimposed on the CT of a COVID positive diagnostic.

Due to the ease of use of these techniques, even with models trained from Transfer Learning, the Grad-CAM is helpful for a subjective analysis of these networks' diagnostic imaging localization capability (PANWAR *et al.*, 2020).

3 Deep Learning

This section will address the primary organization of a Machine Learning (ML) algorithm, specifically those classified within the field of study of Deep Learning. In order to understand how pattern recognition occurs in images using these algorithms, we initially understand the origin of the analogy between biological visual receptors and computational implementations.

Later, we will describe the Perceptrons that constitute the basic mathematical abstraction of ML algorithms and Artificial Neural Networks. Finally, we will explain how we can optimize the learning of these networks so that we can obtain models with accessible computational resources.

3.1 Biological analogy

Since the history of scientific development has been inspired by the ready-made solutions that nature has already produced, we naturally look to neurons for a basic structure with which to build intelligent systems.

Biological neurons are cells highly specialized in receiving and transmitting electrical signals. Moreover, despite the complex tasks possible to be performed with our brain, for example, each of these cells has a straightforward structure.

Each neuron comprises a cell body containing its genetic material, several tiny nerve endings called dendrites, and a more extended nerve ending called an axon. However, it is already possible to highlight other parts of this basic structure, as shown in Figure 3.1

Despite the simple organization of each of these cells, our brain can perform complex tasks because of the efficient communication network of these neurons. The composition of small electrical pulses allows the desired synchronization of many of our biological systems.

With specific regard to our ability to perceive the environment, 1981 Nobel Prize winners David H. Hubel and Torsten Wiesel have developed a series of studies about how living things can understand visual patterns. After experiments in primates, they

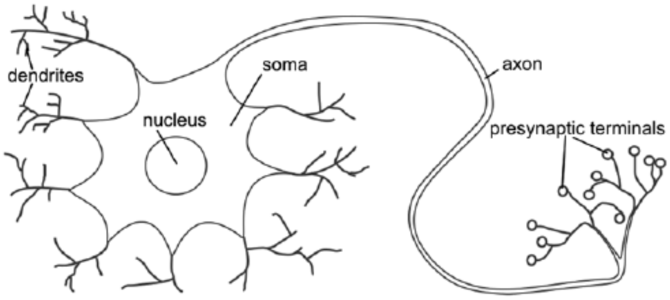

FIGURE 3.1 – Typical neurons consist of a cell body, dendrites, which are fine, one axon, and presynaptic terminals (FRANZE; GUCK, 2010)

were able to identify initially that a portion of our brain has neurons that respond from elementary patterns in the environment, such as lines.

However, they have identified neurons that respond by more complex patterns, such as geometric shapes and sets of curves. And so on, so they also found sets of neurons that are fired only from symbols with intrinsic meaning. They proposed that upper-layer neurons, as seen in Figure 3.2, are stimulated due to their synaptic connections with those from lower layers.

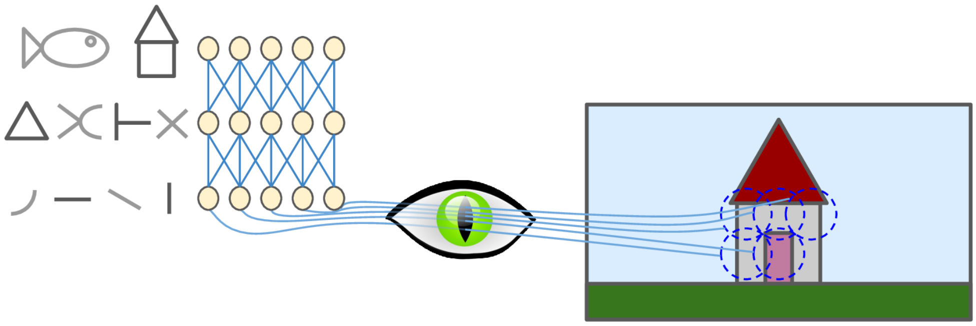

FIGURE 3.2 – Schematic diagram of the hierarchical organization of neurons proposed by David H. Hubel and Torsten Wiesel. Extract from Geron (2019).

This hierarchical organization served as inspiration for constructing the Neocognitron model that simulates a digital computer (FUKUSHIMA *et al.*, 1983). To better understand this biological analogy in digital systems, it is first necessary to understand how we represent these cells in a mathematical structure and how this network-shaped learn patterns.

3.2 Perceptrons and Neural Networks

Despite the popularity of the term Perceptron in the current scientific community, this was not the first artificial neuron model proposed in the literature. In 1943 Warren McCulloch and Walter Pitts proposed that a network of these neurons could learn complex tasks through propositional logic (FITCH, 1944).

In Figure 3.3, assuming that each neuron C lets the signal pass by itself since it receives two simultaneous signals as input, it is possible to describe each logical operation from these neurons' simple organization in different networks.

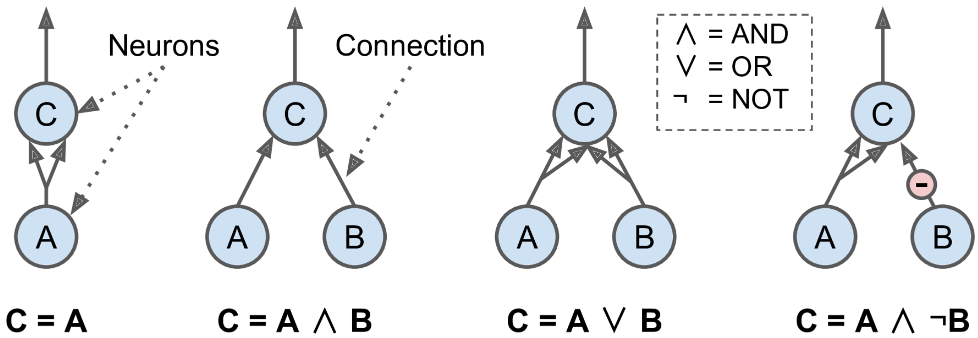

FIGURE 3.3 – A logic representation of neurons according to McCulloch and Pitts'. The main idea is that distinct organizations of neurons can represent logical propositions. Extract from Geron (2019).

However, despite the innovation brought to academia by these researchers, this structure is not the most popular in artificial neural networks used today. Instead of these, the basic mathematical structure of widespread networks today is the Perceptron, first proposed by Frank Rosenblatt in 1957 (ROSENBLATT, 1957).

Perceptrons are similar to Neurons proposed by McCulloch and Pitts but more general. As shown in Figure 3.4, this structure adds weights associated with each of the input signals and a function applied to the linear combination of these signals composing the output of the Perceptron.

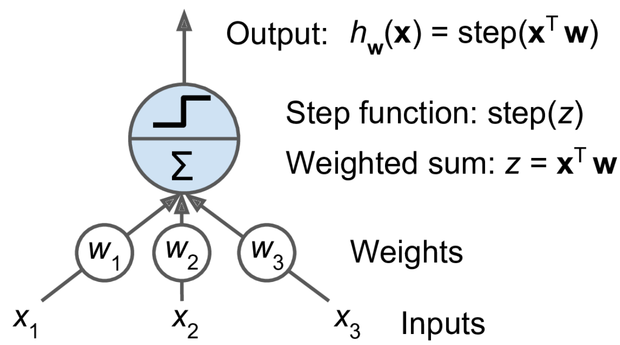

FIGURE 3.4 – X-Ray sample for training process. Extract from Geron (2019)

Mathematically the operation performed by a neuron can be defined as

$$h(z) = h(w_1x_1 + w_2x_2 + w_3x_3 + ... + w_kx_k), \tag{3.1}$$

where a neuron has k inputs. The h function in this case is the Heaviside step function, and historically it has been tried out according to the biological analogy of neurological action potentials (DAYAN, 2001).

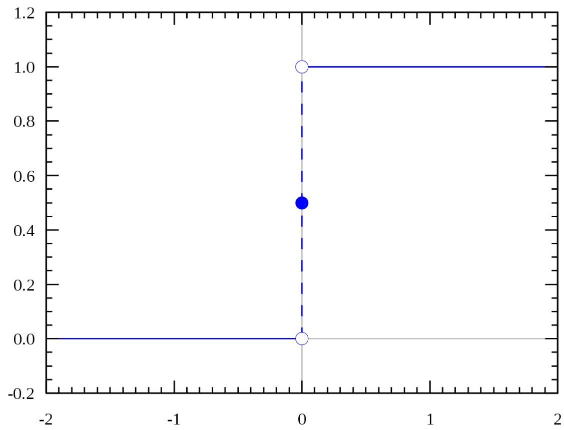

FIGURE 3.5 – Heaviside step function with the half-maximum convention. Extract from Geron (2019)

Like their predecessors, these neurons are commonly networked so that the output of a given neuron makes up a portion of the input of the next neuron. One form of organization is a fully connected type, also known as a dense network. In this type of network, all possible combinations between each of the layers of neurons are carried out.

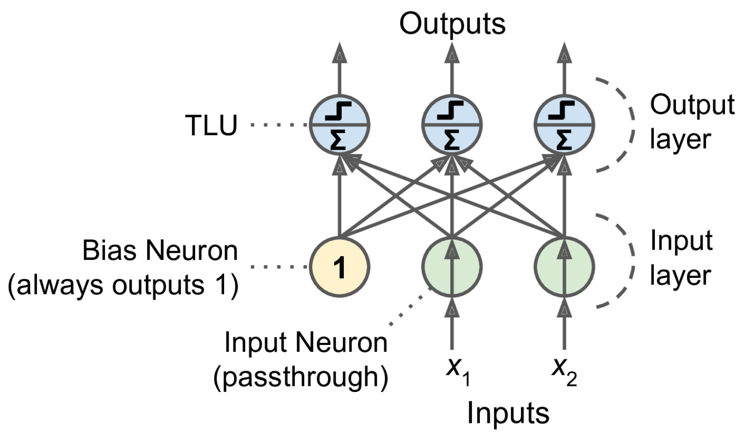

FIGURE 3.6 – X-Ray sample for training process. Extract from Geron (2019)

As we can see in Figure 3.6, in addition to the terms already mentioned, dense neural networks are also composed of constant input signals in each neuron: the bias. This input signal plays a vital role in expanding the learning capacity of networks.

Let us take an example of a function of type $y = mx$ compared to another function of type $y' = mx+c$. It is possible to understand the intuition of the extended generalizability of the bias if we suppose that for the function y, we only had control of the variable m. While for the function y', we had control of the variables m and c.

As shown in Figures 3.7 and 3.8 for the first case, our ability to generalize the y function only includes functions that pass through the origin of the coordinate axis. That

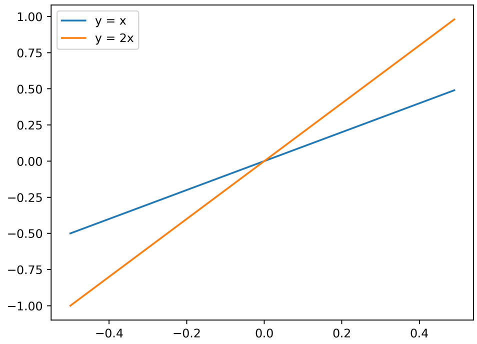

FIGURE 3.7 – Example of two functions according to the assumption defined for function y.

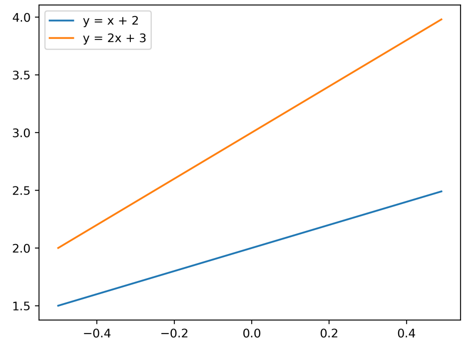

FIGURE 3.8 – Example of two functions according to the assumption defined for function y'.

is only one degree of freedom. For y' function, we represented a more extensive set of functions in the Cartesian plane.

The objective of training the neural networks is to adjust all the network weights to bring the final output closer to the classification objective according to data available for training. When we refer to learning the network, we refer to setting all the weights needed to define the network.

Let θ be this set of all weights and biases intrinsic to the neural network:

$$\theta = \begin{bmatrix} w_1 & w_2 & \dots & w_k & b_1 & b_2 & \dots & b_u \end{bmatrix}^T, \tag{3.2}$$

where k and u depend on the neural network architecture.

Furthermore, it is necessary to define a way to measure whether the neural network's final output is as expected. For this, we need to define a loss function $L(\hat{y}^{(i)}, y^{(i)})$ which lists each of the outputs of the network $y^{(i)}$ and its corresponding expected value $\hat{y}^{(i)}$ according to the training dataset.

A set of cost functions can be used for training neural networks. In our work, we use the cross-entropy function according to the equation

$$L(\hat{y}, y) = ylog(\hat{y}) + (1 - y)log(1 - \hat{y}), \tag{3.3}$$

and which behaves as shown in Figure 3.9.

Intuitively, the idea of the cross-entropy function is to return values as more significant as the error obtained by the network concerning the expected value of the classification. In this sense, the present work uses the supervised learning technique since all training samples previously contain the expected value information.

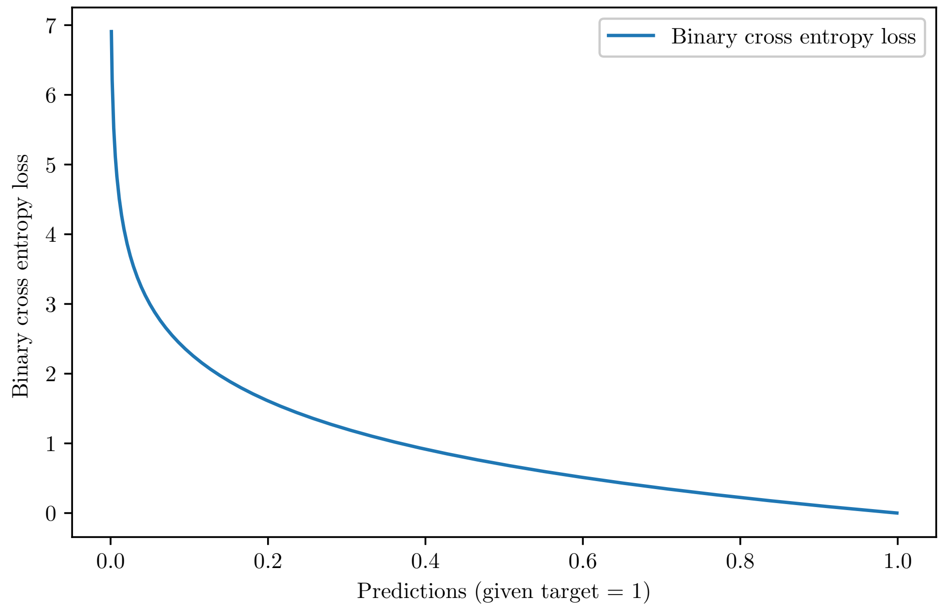

FIGURE 3.9 – The binary cross-entropy loss function for a given target.

Still focusing on a single neuron, to develop learning during training, weights and bias are updated at each iteration n of the algorithm according to the equation

$$\Theta_{n+1} = \Theta_n - \alpha \frac{\partial J(\Theta_n)}{\partial \Theta_n}, \tag{3.4}$$

$$J = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}), \tag{3.5}$$

and J the calculated cost function where α being the learning rate (hyperparameter), $\hat{y}^{(i)}$ is expected value for the $i - th$ output, and m the total number of training samples.

With this mathematical representation, we migrate the learning problem abstractly to a mathematical problem of minimizing the function J. The mathematical representation described above for the training of Perceptrons was initially proposed by Frank Rosenblatt, inspired by Donald Hebb's book "The Organization of Behavior" (SHAW, 1986).

Briefly, the intuition behind this algorithm is that when a given neuron sends a signal of significant strength to another, the connection between them must become more robust. That is, the cost of prediction must decrease.

As we have seen, a set of interspersed layers of perceptrons define an Artificial Neural Network (ANN). From this definition, neural networks are classified as deep once there are many layers. Therefore, we were able to define and understand the meaning behind the term Deep Learning. In other words: the process of learning by Deep Neural Networks.

We realize that the amount of weights and bias needed to define the architecture of a given neural network grows exponentially with the number of neurons in the network if we assume the idea of a fully connected network.

Thus, given this considerable amount of parameters to be defined, for a long time, the scientific community searched without success for viable techniques for learning these networks to be applied in complex problems.

David Rumelhart, Geoffrey Hinton, and Ronald Williams solved this problem by publishing the seminal paper: Learning Internal Representations by Error Propagation (RUMELHART; MCCLELLAND, 1987).

The idea they proposed and used until today for learning networks is to introduce the concept of a backpropagation training algorithm. This technique allows the calculation of the necessary gradients in the step of learning the equation 3.4 very efficiently through matrix operations.

3.3 Convolutional Neural Networks

Even with the optimizations found to guarantee the learning of Deep Neural Networks, this type of architecture is not the most used in tasks related to computer vision.

Based on the biological analogy of the functioning of our visual cortex, a group of researchers proposed architectures in the form of convolutional neural networks. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, who introduced the famous LeNet-5 architecture in 1998, proposed two ideas for optimizing the learning of neural networks focused on image recognition: The pooling layers and the convolutional layers (LECUN *et al.*, 1998).

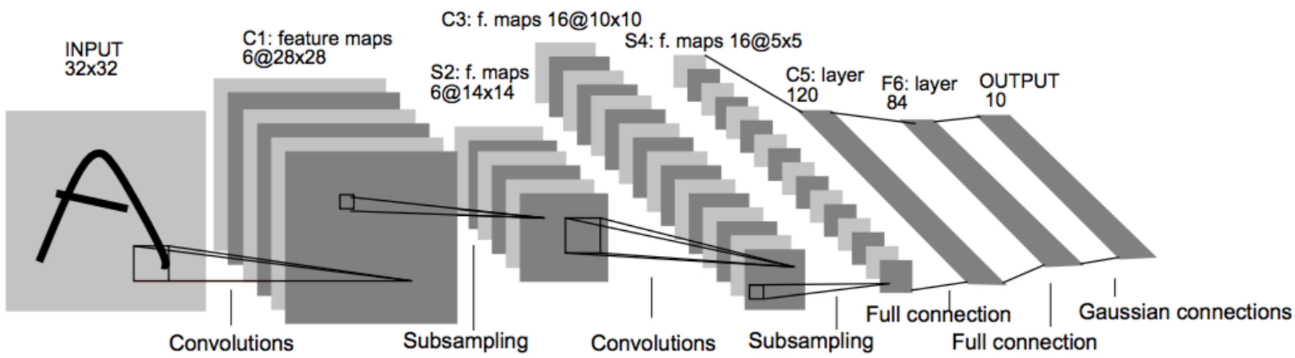

FIGURE 3.10 – The LeNet-5 architecture: A Convolutional Neural Network. Extract from Lecun *et al.* (1998)

3.3.1 Pooling layers

Pooling layers can be of different types. Figure 3.12 presents the idea of this type of operation. For a given square matrix of fixed size, the max-pooling operation is an operation that chooses the maximum value of that matrix. So, since an operation like this is performed for an entire image, matrix by matrix, we have a reduced image since we reduced the resolution of the original input image.

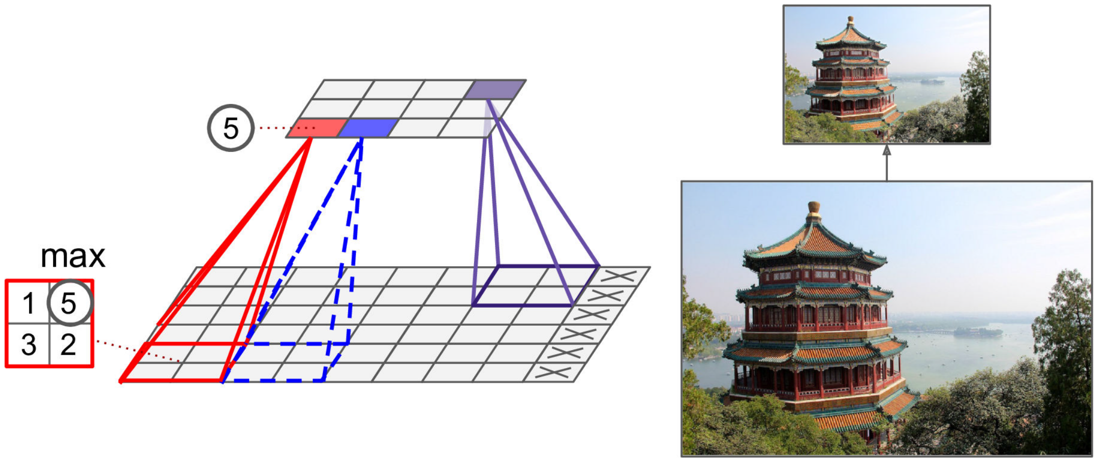

FIGURE 3.11 – X-Ray sample for training process. Extract from Geron (2019)

As we can see in Figure 3.12, let three different images, A, B, and C be input for a max-pooling operation. The images obtained as the output of inputs A and B are identical. We can see that this type of operation applies a certain invariance over small input layer translations.

In tasks requiring some mastery in learning to describe environments, slight variations between one image and another do not necessarily reflect in different classifications.

3.3.2 Convolutional layers

Convolutional layers work from the idea of receptive fields. Unlike fully connected layers, which make up the architecture of dense networks, each of the neurons in a convolutional layer does not connect with the previous layer.

The set of neurons of a given layer that connect with a given neuron of a posterior layer makes up precisely the receptive field of the latter. Figure 3.14 highlights the receptive field of a convolutional second-layer neuron as well as two first-layer neurons.

A question that may arise is what is the meaning associated with each of these convolutional layers. This meaning depends on the convolution operation to be applied. In practical terms, a convolutional layer results from a convolution operation applied to the previous layer. The convolution operation type is also known as Convolutional Filter.

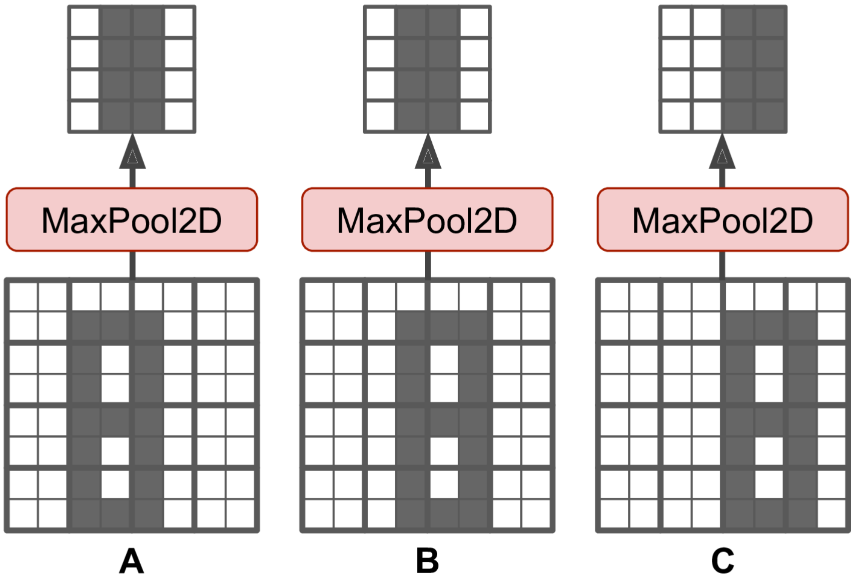

FIGURE 3.12 – Schematic representation of a 2D Max Pooling operation. Extract from Geron (2019).

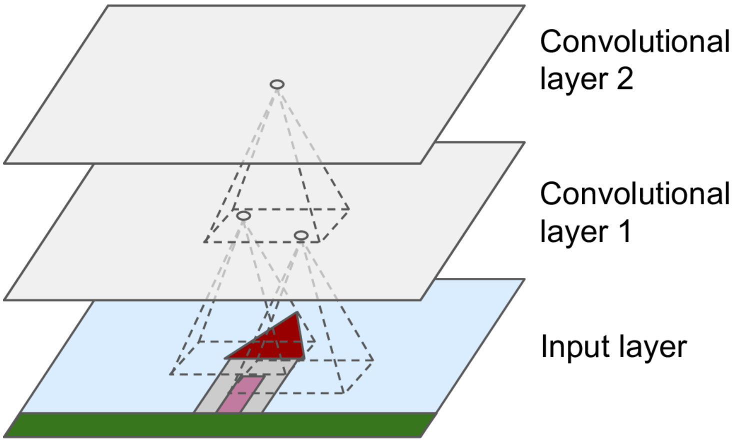

FIGURE 3.13 – Schematic representation of two 2D convolution operations. Extract from Geron (2019).

Once the result of a convolution operation to a given pixel matrix that describes a channel of an image, the output of that operation is another pixel matrix. The characteristics highlighted in the latter layer depends on the properties of the convolutional filter used, as we can see in Figure 3.14.

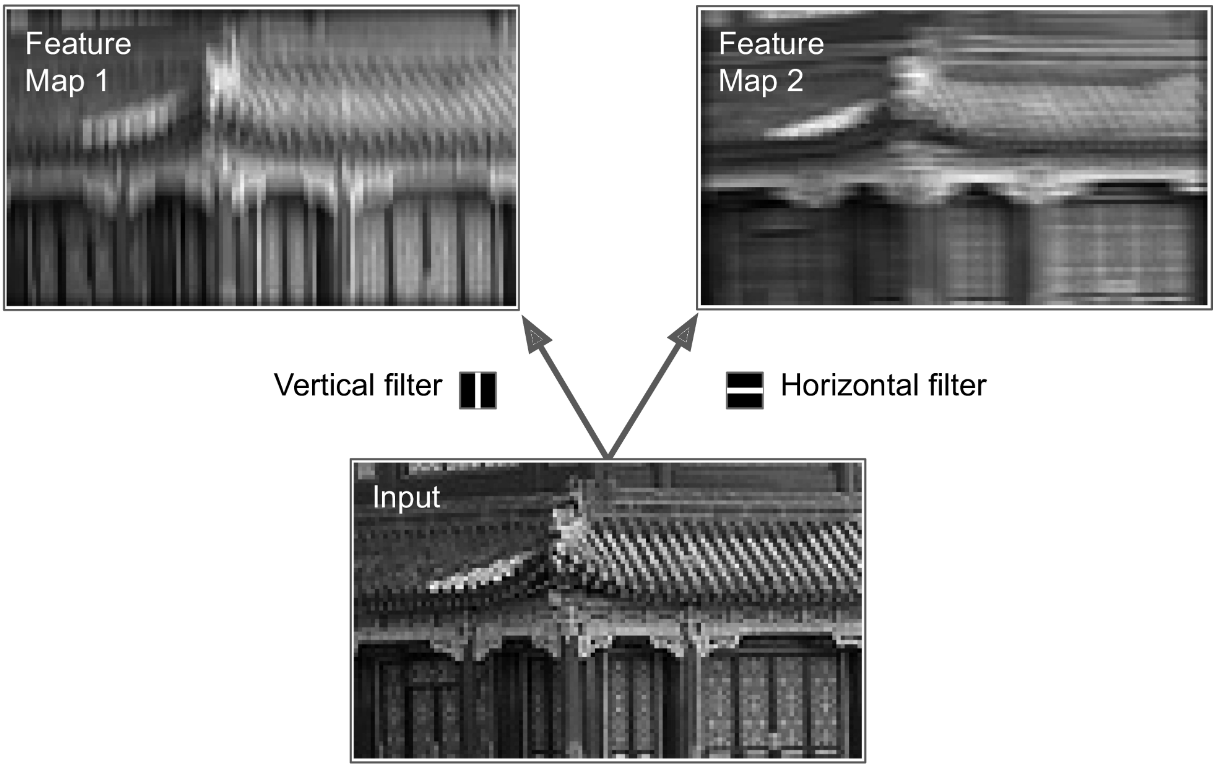

FIGURE 3.14 – Schematic representation of two distinct convolution operations applied to the same image. Extract from Geron (2019).

4 Methodology

This chapter consists of a description of the methodology used during the experiments. We chose the dataset and the final architecture of the network through several tests until we found the most satisfactory results possible.

The reproducibility of the results obtained is directly related to the reproducibility of the methodology presented here. All Python code was developed using the Tensorflow 2.0 framework.

4.1 Dataset

Even though it is one of the worst health crises in the world in recent decades, it is still challenging to find good sets of medical data that allow for more significant discussion in the scientific environment.

After an extended exploration, we found a work that proposed to build a Covid classification model using computed tomography images, just as we did. This work also provided a dataset resulting from a set of 7 other similar datasets (MAFTOUNI *et al.*, 2021).

The authors of that work guaranteed that the data set did not have repeated images. In addition, a set of extra information about the patients was made available, such as the country of origin, gender, and age of the patients, as can be seen in Figure 4.1 (MAFTOUNI *et al.*, 2021).

Initially, we performed an adequate data segmentation between 2 sets of data for training and analysis of the performance of our model. The training and validation datasets were set up so that a patient is not in both datasets, which could lead to training overfitting. This could occur considering that in the case of CT scans, as opposed to X-Ray images, the same patient may have different lung segmentation, as shown in Figure 4.4.

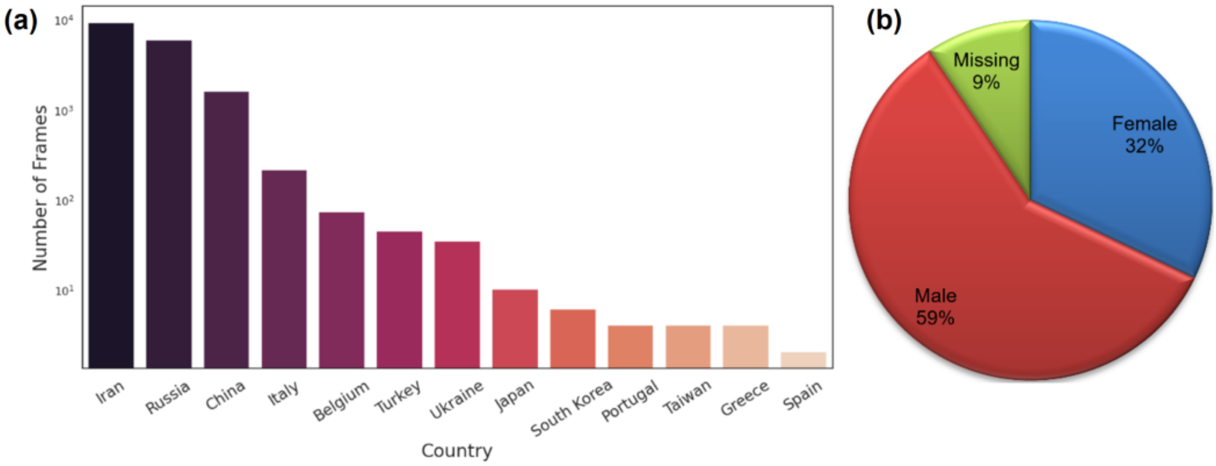

FIGURE 4.1 – Information about the dataset used for training and validation of models. (a) the number of images per country of origin of patients. (b) the gender distribution among patients. Extract from Maftouni *et al.* (2021).

4.2 Neural Network architecture

The base architecture of the neural network used during the experiments was the InceptionResNetV2 (SZEGEDY *et al.*, 2016). This architecture combines the ideas of neural networks of residual connections with neural networks of Inception architecture, as we can see in Figure 4.2, which presents a simplified diagram of the network.

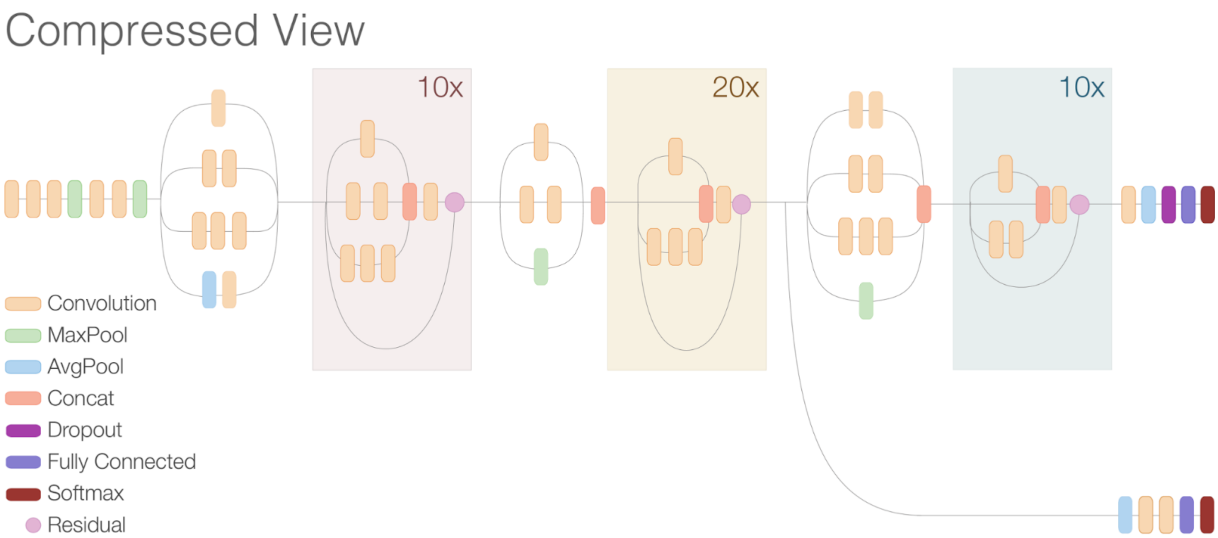

FIGURE 4.2 – Representation of the architecture of the InceptionResnetV2 network. It is possible to see the presence of convolutional layers and MaxPool as already mentioned in this work. Extracted from Google... (2021)

As will be detailed in the training section, it was necessary to add more layers of neurons on top of the network to optimize our training. The added layers were of fully wired neurons, as shown in Figure 4.3.

More details about the implementation of the neural network and the experiments that culminated in the chosen architecture are available at: `https://github.com/ArthurMor4is/grad-cam-covid-19-ct`.

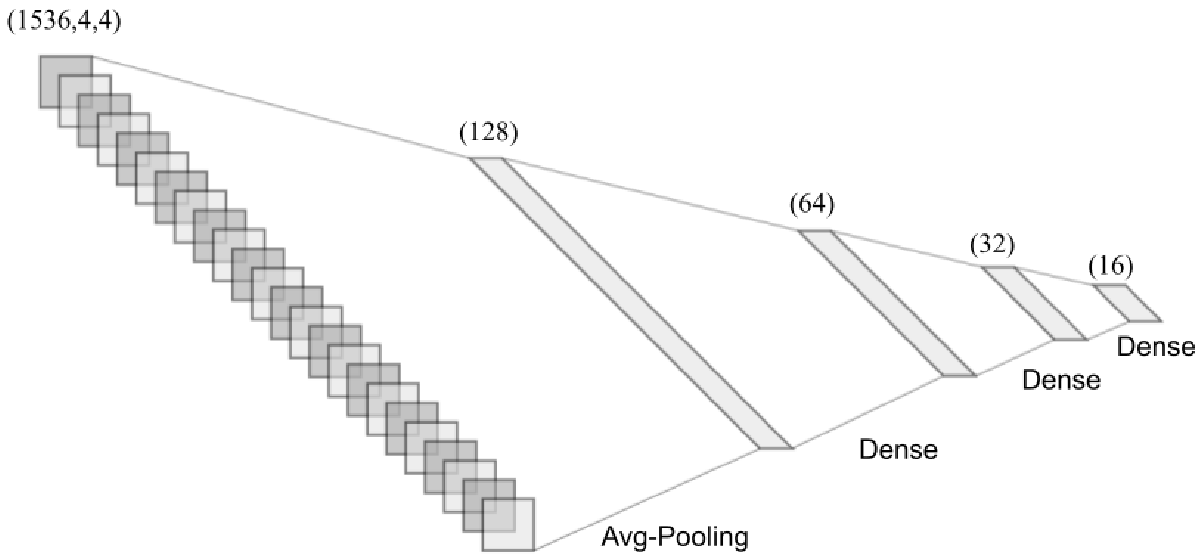

FIGURE 4.3 – Diagram representing the layers of neurons added to the pre-trained InceptionResnetV2 network. These new layers from the start have randomly initialized weights. The last layer of this network was omitted, but it is only a single neuron as the output of the COVID classification task.

4.3 Training procedure

The training of the neural network used the Transfer Learning strategy following the steps:

1. Base model instantiated and loaded pre-trained weights into it.
2. All layers are frozen in the base model by setting `trainable` as `False`.
3. A New model is created on top of the output of one (or several) layers from the base model.
4. The new model is trained on your new dataset.

In addition to this basic step-by-step, a Fine-tuning strategy was also used to improve training performance. To learn more, see Chollet (2020).

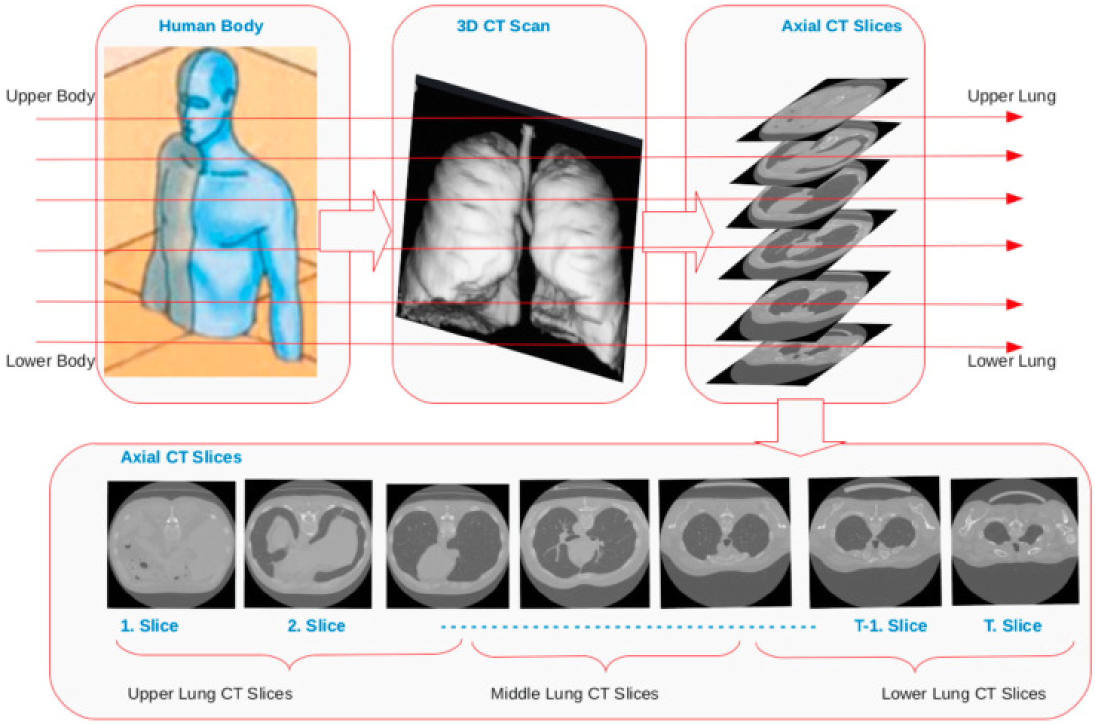

FIGURE 4.4 – Schematic representation of a computed tomography exam. A single patient can be represented by a large set of images of their lung.

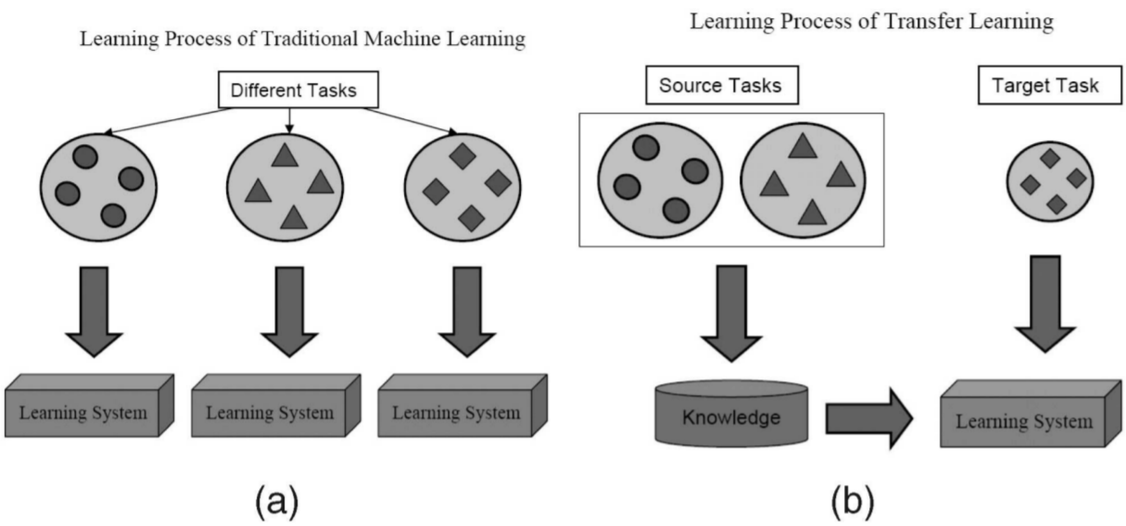

FIGURE 4.5 – (a) The traditional machine learning algorithm learning process.(b) The process required to use the Transfer Learning strategy.

5 Results and discussions

As explained in the training section, due to the use of the transfer learning method with fine-tuning, the training was divided into two parts. For the first part, in which the layers of the InceptionResnetV2 network are frozen, we will call it the First step model.

That is, this denomination refers to the model trained at the end of this step. In other words, the weights of this model for the layers of the InceptionResnetV2 network are the same that this network acquired when being trained in the ImageNet image set. And so, the training of this step only updates the weights of the additional dense layers.

Similarly, as the next step of training is to update all weights in the net, in this section, the name "Second step model" refers to the final model after the last step of training, that is, updating all weights.

Figures 5.6, 5.5, 5.7, and 5.8 present the behavior of these models over the training periods, and table 5.1 presents the final performance of the models. Due to the use of the Early-Stopping technique, it is possible to see that the metrics presented in the table differ from the final metrics in the graphs. This fact is due to the use of a Rollback strategy.

That is, at a certain point during training, it is noticed that the metrics obtained between the training and validation datasets start to diverge, the training is paused, and the weights of the best results so far are that they are saved in the final model.

As expected, when applying Grad-CAM on different performance networks, we got different heatmaps. As seen in Figures 5.1, 5.2, 5.3, 5.4 the heat maps obtained for the network with the best performance are more located in the regions close to the lung.

For each of the heatmaps, our implementation assigns a color scale proportional to the intensity of the feature map processed by Grad-CAM. The colors more to the blue spectrum indicate less significant regions for the decision-making of the network. While regions more to the red of the spectrum indicate the more significant regions.

	Accuracy	Loss
First step model	68.3%	33.6%
Second step model	92.8%	22.8%

TABLE 5.1 – Accuracy and loss information for both analyzed models. The `threshold` for accuracy was `0.5`.

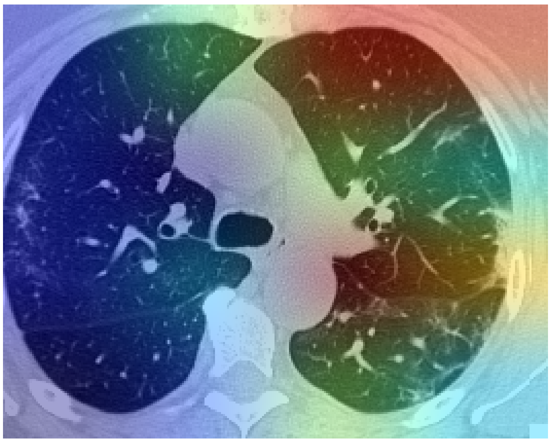

FIGURE 5.1 – Heat map obtained after the first training. It is noticeable that the heat map generated indicates regions outside the limits of the lungs

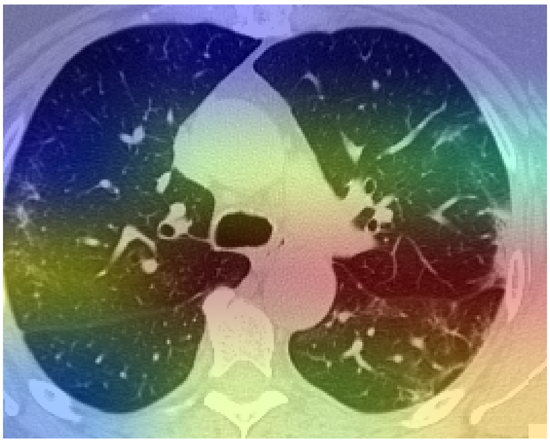

FIGURE 5.2 – Heat map obtained after the first training. With a more accurate network, the identification of patterns in the lung would be expected, as seen in the figure.

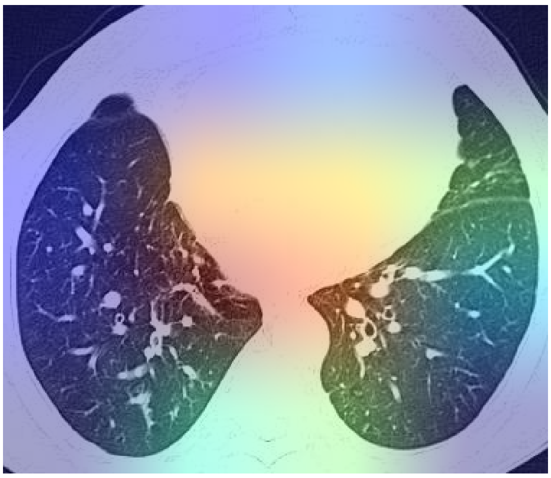

FIGURE 5.3 – Heat map obtained after the first training of the model. The highlighting of regions between the lungs is noticeable, which at first would not make sense for the diagnosis decision.

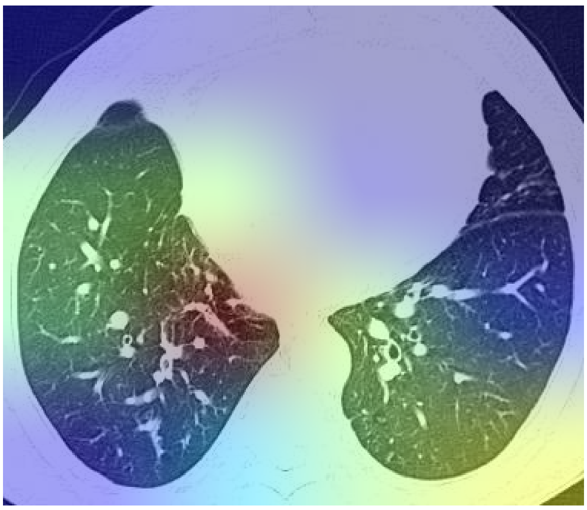

FIGURE 5.4 – Heat map obtained after the end of training. The model was able to highlight features closer to the patient's left lung and did not show a red color in the center of the image.

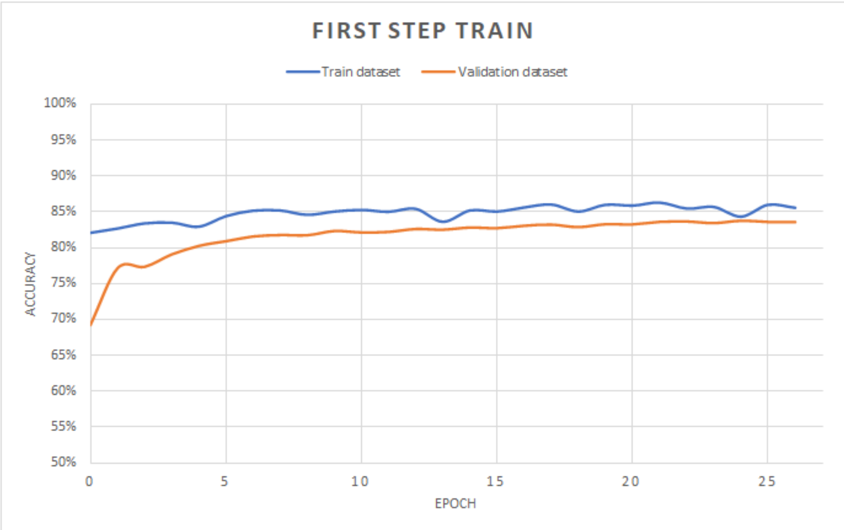

FIGURE 5.5 – Time series of model accuracy during the first training step.

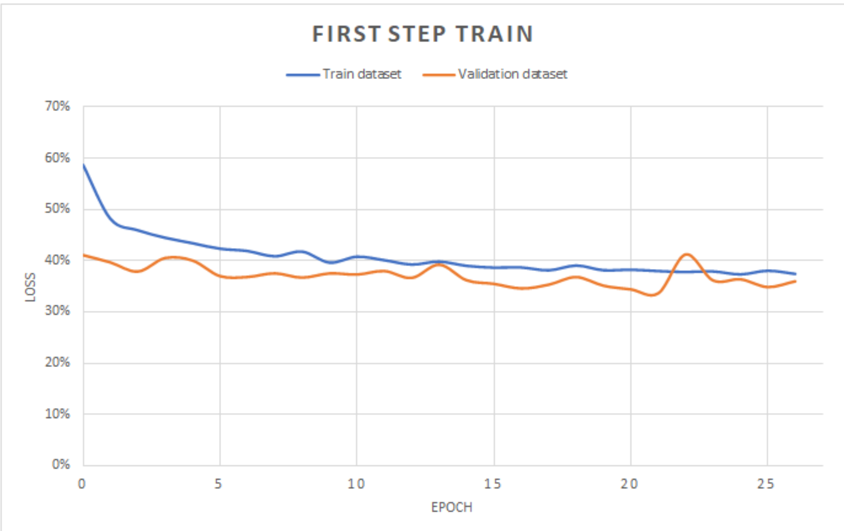

FIGURE 5.6 – Time series of model loss during the first training step.

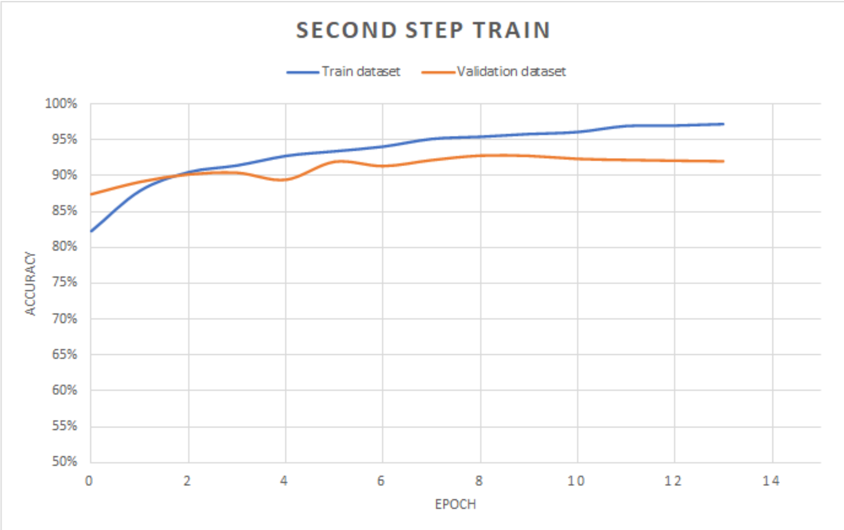

FIGURE 5.7 – Time series of model accuracy during the second training step.

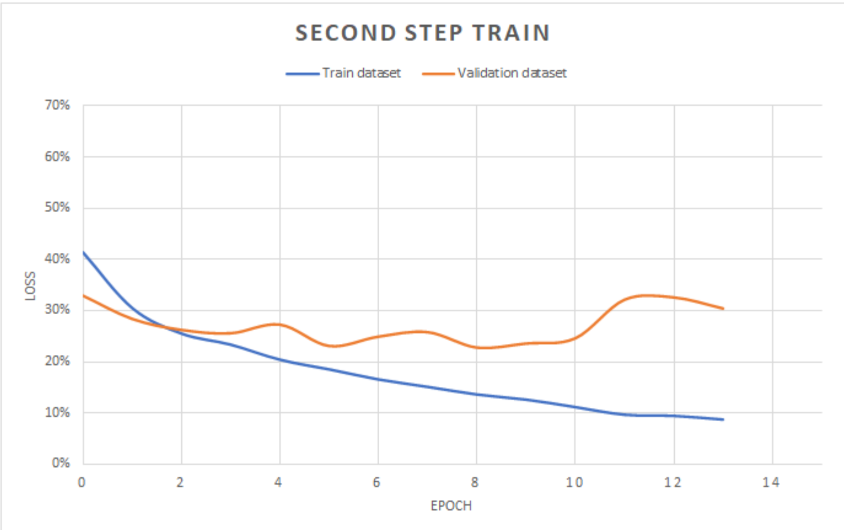

FIGURE 5.8 – Time series of model loss during the second training step.

6 Conclusion

It is undeniable that the field of artificial intelligence has developed substantially in recent decades. The industry has used increasingly customizable and autonomous solutions, and in parallel, the scientific community today has more resources available to accelerate the experimentation cycle.

However, this optimism bumps into problems that, if not resolved, can undesirably impact the expected efficiency of models in natural environments. In this context, the sphere of medicine may be the most plausible sector for us to have more significant investments since, as we have seen, it has a great potential to impact the lives of millions of people.

Society has evolved substantially in increasingly surprising technologies and at an ever-increasing speed. Nevertheless, we have failed once again to organize ourselves to control a worldwide epidemic that has resulted in the death of millions of people, and many of them died needlessly.

It is also undeniable that the field of study of techniques that seek to explain the decision-making of supervised-trained algorithms is still very new. However, given the mentioned facts, it is for this reason that the study of these techniques is motivating, as they can aid in the development of increasingly reliable algorithms.

The Grad-CAM technique proved to be efficient in the present work for two models with different performances. The technique proved to be effective in generating outputs with different meanings. This fact does not imply concluding that the technique proves to be effective as a necessary and sufficient condition for analyzing the performance of previously trained models.

Given that the generation of heat maps by Grad-CAM takes place from the last convolutional layer of the trained networks, this layer often has a much lower resolution than the images that served as training for the models. For example, in our case, the last convolutional layer was a 16-pixel array.

Therefore, one must pay attention to the fact that the superposition of this generated map, with the initial image that served as training, may imply vague interpretations of the model's behavior. One pixel highlighted in the heatmap will be associated with several pixels in the original training image.

Thus, Grad-CAM proved to be efficient in generating heat maps and, in fact, worked differently for different models. This fact indicates that it served well for developing an artificial neural network concerning validating the trained model.

6.1 Future works

A possible extension of this work is to use 3D analysis through Grad-CAM or in parallel with other explainability techniques. In addition, we also intend to experiment with other possible color spectra to define the behavior of maps. It is also possible to use explicability techniques outside the domain of medicine since this type of procedure does not change the behavior of previously trained networks. We also intend to share with the scientific community more results from this type of application.

Bibliography

ADEBAYO, J.; GILMER, J.; MUELLY, M.; GOODFELLOW, I. J.; HARDT, M.; KIM, B. Sanity checks for saliency maps. **CoRR**, abs/1810.03292, 2018. Available at: http://arxiv.org/abs/1810.03292.

CHOLLET, F. **Keras documentation: Transfer learning fine-tuning**. François Chollet, Apr 2020. Available at: https://keras.io/guides/transfer_learning/.

DATA, O. W. in. **Coronavirus (COVID-19) Cases**. 2021. Available at: https://ourworldindata.org/covid-cases.

DAYAN, P. **Theoretical neuroscience : computational and mathematical modeling of neural systems**. Cambridge, Mass: Massachusetts Institute of Technology Press, 2001. ISBN 978-0262541855.

FITCH, F. B. Warren s. mcculloch and walter pitts. a logical calculus of the ideas immanent in nervous activity. bulletin of mathematical biophysics, vol. 5 (1943), pp. 115–133. **Journal of Symbolic Logic**, Cambridge University Press, v. 9, n. 2, p. 49–50, 1944.

FRANZE, K.; GUCK, J. The biophysics of neuronal growth. **Reports on Progress in Physics**, v. 73, p. 094601, 08 2010.

FUCHS, T. Ethical issues in neuroscience. Ovid Technologies (Wolters Kluwer Health), v. 19, n. 6, p. 600–607, nov. 2006. Available at: https://doi.org/10.1097/01.yco.0000245752.75879.26.

FUKUSHIMA, K.; MIYAKE, S.; ITO, T. Neocognitron: A neural network model for a mechanism of visual pattern recognition. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-13, n. 5, p. 826–834, 1983.

GARDNER, H. What we do don't know about learning. **Daedalus**, The MIT Press, v. 133, n. 1, p. 5–12, 2004. ISSN 00115266. Available at: http://www.jstor.org/stable/20027891.

GERON, A. **Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. 2. ed. Sebastopol, CA: O'Reilly Media, 2019.

GHASSEMI, M.; OAKDEN-RAYNER, L.; BEAM, A. L. The false hope of current approaches to explainable artificial intelligence in health care. Elsevier BV, v. 3, n. 11, p. e745–e750, nov. 2021. Available at: https://doi.org/10.1016/s2589-7500(21)00208-9.

GOOGLE AI Blog: Improving inception and image classification in tensorflow. *In*: . [*S.l.: s.n.*], 2021. Available at: https://ai.googleblog.com/2016/08/improving-inception-and-image.html. Accessed on: 11 nov. 2021.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. MIT Press - Journals, v. 18, n. 7, p. 1527–1554, jul. 2006. Available at: https://doi.org/10.1162/neco.2006.18.7.1527.

KOOMEY, J.; BERARD, S.; SANCHEZ, M.; WONG, H. Implications of historical trends in the electrical efficiency of computing. Institute of Electrical and Electronics Engineers (IEEE), v. 33, n. 3, p. 46–54, mar. 2011. Available at: https://doi.org/10.1109/mahc.2010.28.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, p. 2278 – 2324, 12 1998.

LEE, E.; ZHENG, J.; COLAK, E.; MOHAMMADZADEH, M.; HOUSHMAND, G.; BEVINS, N.; KITAMURA, F.; ALTINMAKAS, E.; REIS, E.; KIM, J.-K.; KLOCHKO, C.; HAN, M.; MORADIAN, S.; MOHAMMADZADEH, A.; SHARIFIAN, H.; HASHEMI, H.; FIROUZNIA, K.; GHANAATI, H.; GITY, M.; YEOM, K. Deep covid detect: an international experience on covid-19 lung detection and prognosis using chest ct. **npj Digital Medicine**, v. 4, 01 2021.

MAFTOUNI, M.; LAW, A. C. C.; SHEN, B.; ZHOU, Y.; YAZDI, N. A.; KONG, Z. A robust ensemble-deep learning model for covid-19 diagnosis based on an integrated ct scan images database. *In*: . **Proceedings** [...]. [*S.l.: s.n.*], 2021.

MINAR, M. R.; NAHER, J.; AL et. Recent advances in deep learning: An overview.

PANWAR, H.; GUPTA, P.; SIDDIQUI, M. K.; MORALES-MENENDEZ, R.; BHARDWAJ, P.; SINGH, V. A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. **Chaos, Solitons Fractals**, v. 140, p. 110190, 2020. ISSN 0960-0779. Available at: https://www.sciencedirect.com/science/article/pii/S0960077920305865.

ROSENBLATT, F. **The perceptron - A perceiving and recognizing automaton**. Ithaca, New York, January 1957.

RUMELHART, D. E.; MCCLELLAND, J. L. Learning internal representations by error propagation. *In*: ______. **Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations**. [*S.l.: s.n.*], 1987. p. 318–362.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. **ImageNet Large Scale Visual Recognition Challenge**. 2015.

SELVARAJU, R. R.; DAS, A.; VEDANTAM, R.; COGSWELL, M.; PARIKH, D.; BATRA, D. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. **CoRR**, abs/1610.02391, 2016. Available at: http://arxiv.org/abs/1610.02391.

SHAW, G. L. Donald hebb: The organization of behavior. *In*: . Springer Berlin Heidelberg, 1986. p. 231–233. Available at: https://doi.org/10.1007/978-3-642-70911-1_15.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. **Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning**. 2016.

UN: United nations. *In*: . [*S.l.: s.n.*], 2020. Available at: https://www.un.org/en/desa-/covid-19-slash-global-economic-output-85-trillion-over-next-two-years. Accessed on: 7 jun. 2021.

VILONE, G.; LONGO, L. **Explainable Artificial Intelligence: a Systematic Review**. 2020.

WANG, M.; XIA, C.; HUANG, L.; XU, S.; QIN, C.; LIU, J.; CAO, Y.; YU, P.; ZHU, T.; ZHU, H.; WU, C.; ZHANG, R.; CHEN, X.; WANG, J.; DU, G.; ZHANG, C.; WANG, S.; CHEN, K.; LIU, Z.; XIA, L.; WANG, W. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. Elsevier BV, v. 2, n. 10, p. e506–e515, out. 2020. Available at: https://doi.org/10.1016/s2589-7500(20)30199-0.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO N°	4. N° DE PÁGINAS
TC	25 de novembro de 2021	DCTA/ITA/TC-132/2021	46

5. TÍTULO E SUBTÍTULO:

HEATMAP ANALYSIS IN COVID-19 CT IMAGES BASED ON CONVOLUTIONAL NEURAL NETWORKS

6. AUTOR(ES):

Arthur Fernandes de Morais

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica – ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

COVID, Convolution, GRAD-CAM, Explainability, Neural

9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Intelegência artificial: Aprendizagem (inteligencia artificial); Redes neurais; SARS-CoV2; Diagnosticos; Tomada de decisão; Medicina; Computação.

10. APRESENTAÇÃO: (**X**) **Nacional** ( ) **Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia de Computação. Orientador: Marcus Henrique Victor Júnior; co-orientadores: Mônica Mitiko Soares Matsumoto e Marcos Ricardo Omena de Albuquerque Maximo. Publicado em 2021.

11. RESUMO:

During the last few decades, we have supported the hypothesis that our method of logical reasoning is justifiably accurate in developing computational algorithms analogous to the workings of our intellect. Thus, it is natural to doubt whether these same algorithms are making decisions consistent with the environment with this same logic. The need to obtain plausible justifications for a given decision is as more significant as the risk associated with that decision. Allied to this, we have seen more and more computer applications inserted in medical environments due to the benefits of using artificial intelligence to reduce clinical errors. Even though we have increased human life expectancy worldwide, we are still subject to global pandemics that primarily affect the poor, as became apparent during the COVID-19 pandemic that started in 2019. Inspired by this entire context, this work proposes to explore an explicability technique that does not depend on the artificial neural network architecture used in machine learning training. More specifically, through the Grad-CAM technique, it was possible to analyze the decision-making of a neural network for the diagnosis of SARS-COV-2. The developed neural network obtained 92.8% accuracy in diagnosing COVID-19 from 13,993 computed tomography scans of lungs from patients from different countries. Despite the scope of the work, it should be noted that the technique used can be applied in other fields beyond medicine.

12. GRAU DE SIGILO:

(**X**) **OSTENSIVO** ( ) **RESERVADO** ( ) **SECRETO**