COVID-19 automatic diagnosis with CT images using the novel Transformer architecture

Gabriel Sousa Silva Costa[1], **Anselmo C. Paiva**[1], **Geraldo Braz Junior**[1], **Marco Melo Ferreira**[2]

Núcleo de computação aplicada – Universidade Federal do Maranhão (UFMA)
65.080-805 – São Luís – MA – Brasil[1]

Rodovia MA 201, Km 12, s/n - Piçarreira, São José de Ribamar - MA, BR[2]

Abstract. *Even though vaccines are already in use worldwide, the COVID-19 pandemic is far from over, with some countries re-establishing the lockdown state, the virus has taken over 2 million lives until today, being a serious health issue. Although real-time reverse transcription-polymerase chain reaction (RT-PCR) is the first tool for COVID-19 diagnosis, its high false-negative rate and low sensitivity might delay accurate diagnosis. Therefore, fast COVID-19 diagnosis and quarantine, combined with effective vaccination plans, is crucial for the pandemic to be over as soon as possible. To that end, we propose an intelligent system to classify computed tomography (CT) of lung images between a normal, pneumonia caused by something other than the coronavirus or pneumonia caused by the coronavirus. This paper aims to evaluate a complete self-attention mechanism with a Transformer network to capture COVID-19 pattern over CT images. This approach has reached the state-of-the-art in multiple NLP problems and just recently is being applied for computer vision tasks. We combine vision transformer and performer (linear attention transformers), and also a modified vision transformer, reaching 96.00% accuracy.*

1. Introduction

Real-time reverse transcription polymerase chain reaction (RT-PCR) is the first choice for COVID-19 diagnosis, but its high false-negative rate and low sensitivity may delay accurate diagnosis [1]. Thus, it is necessary to explore alternative diagnosis methods. Fortunately, the chest's x-ray and CT scans show signals that differentiate a normal healthy thorax from an unhealthy one. With overloaded healthcare systems all around the world, it is important to quickly assess if incoming patients are infected with COVID-19 or not, giving priority to treating patients who are more vulnerable and are infected with the virus. With that in mind, a fast, low-sensitivity and precise diagnosis tool can be very helpful to radiologists to distinguish between patients infected and not infected with the virus. The majority of recently published works applies Convolutional Neural Networks(CNN) architectures [5, 6, 7] for this goal.

Recent works on CT images classification use a tailored deep convolutional network [5] for the problem of classifying medical images. On the same problem, but with chest x-ray, images are also classified using CNNs, using techniques such as capsule based-networks [6] or plain deep convolutional neural networks [7]. A deep features approach with the Q-deformed entropy handcrafted features was evaluated in [13] reaching accuracy of 99.68%. A DenseNet201 with transfer learning was used to train a model

to classify CT chest images between covid and non covid in [14] while comparing with other CNN pretrained models such as VGG16, ResNet152V2 and Inception-ResNetV2 which reach an overall accuracy of 96.25% on the test set. A tailored convolutional neural network, made specifically for COVID-19 detection, reaches 99.1% accuracy on the test set in [5].

We propose a method for COVID-19 diagnosis using CT scans from patients, using complete self-attention networks: Vision Transformer(ViT) [3] and some of its variations(detailed on Section 2.2). Transformers are mainly used for Neural Language Processing problems but we use it as a low-sensitivity and high precision diagnosis tool to classify a given patient chest tomography between a healthy one, pneumonia caused by something other than the corona virus or pneumonia caused by the virus.

The paper is organized as follows. Section 2 describes the proposed methods. Section 3 discusses our findings, followed by Section 4 concludes the work.

2. Proposed Method

In this work, we focused on diagnosis of COVID-19 using CT images with a Transformer Network. For that, we propose a method illustrated on Figure 1, that trains and evaluates with some of the Transformer variants for the image classification problem in order to perform neural architecture estimation.

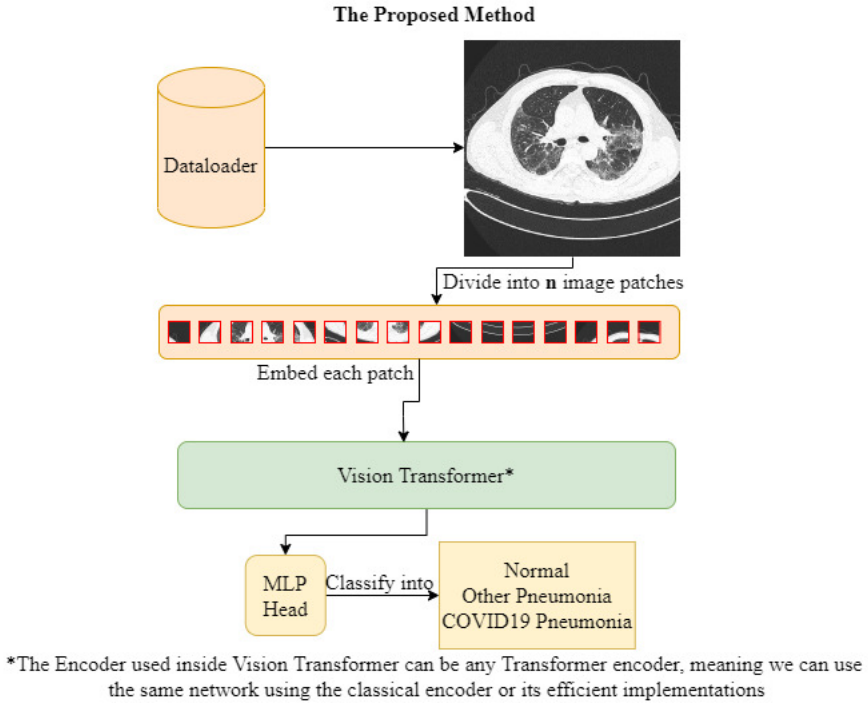

Figure 1. Proposed method

The first step is to preprocess our data. The data is then fed into one of the proposed Vision Transformer variations (Vision Transformer (ViT), Data-efficient Image Transformer (DeIT)). The Encoder's output is then used by the trainable MLP Head, which results in the classification of the image between the three classes we are using to evaluate our model. Each of these steps is explained in following sections.

Table 1. COVIDNet-CT Dataset: Images distribution

Type	Normal	Other Pneumonia	COVID-19 Pneumonia	Total
Training	27201	22061	12520	61782
Validation	9107	7400	4529	21036
Test	9450	7395	4346	21191

2.1. Data Acquisition and Preprocessing

We utilized the COVIDNet-CT [12] dataset that have training, testing and validation sets defined in Table 1.

Our preprocessing consists on loading each set and applying its respective data augmentation: for the training set, we apply a horizontal and vertical random flip, image rotation ranging from -30 to 30 degrees and a random brightness variation, ranging from -0.2 to 0.2. For every set, we apply a resizing according to the network required input we using, (512x512 for ViT and 384x384 for DeIT) and data normalization.

2.2. Vision Transformer

In our work, we use Vision Transformer as a feature extractor to classify CT images between healthy, pneumonia due to infection with the new Corona Virus or pneumonia due to other kinds of infection. We also compare between different implementations of the network.

Vision Transformer(ViT) [3] is a network proposed by Google's Research team, created to assess the classical Transformer's [2] limitation, which is used mainly for Neural Language Processing problems: its prohibitive large memory usage for images. ViT showed promising results on the Image Net dataset, surpassing top-classifying solutions. The big picture of the proposed method can be seen on Figure 1.

ViT's[3] main idea is to make the Transformer architecture feasible for images with minimal modification from the original work [2], so the architecture process an image as if they were a sequence of input tokens, where a fixed-size image is decomposed into N patches of fixed size and each patch is projected with a linear layer, generating embeddings. The embeddings are then concatenated with a class token, a trainable vector that is lately projected with a linear layer to predict each class. Given the fact that a Transformer block is invariant to the order of the patches(a desirable feature for images, given its 2D structures), positional information is incorporated as positional embeddings, which are summed to the embeddings, and then fed to the Transformer blocks. Figure 2 depicts a detailed scheme of the Vision Transformer.

The main ViT concept is its self-attention mechanism, or the more elaborate Multi-head self-attention(MSA). Attention mechanism is based on a trainable associative memory between key and value vector pairs, where a query vector $q \in R^d$ is matched against a set of k key vectors $K \in R^{k \times d}$ using inner products, which are then scaled and normalized with a softmax function to obtain weights. The output is the weighted sum of a set of k value vectors $V \in R^{k \times d}$ for a sequence of N query vectors $Q \in R^{N \times d}$, producing the output matrix of size $N \times d$:

$$Attention(Q, K, V) = Softmax(QK^T/\sqrt{d})V \quad (1)$$

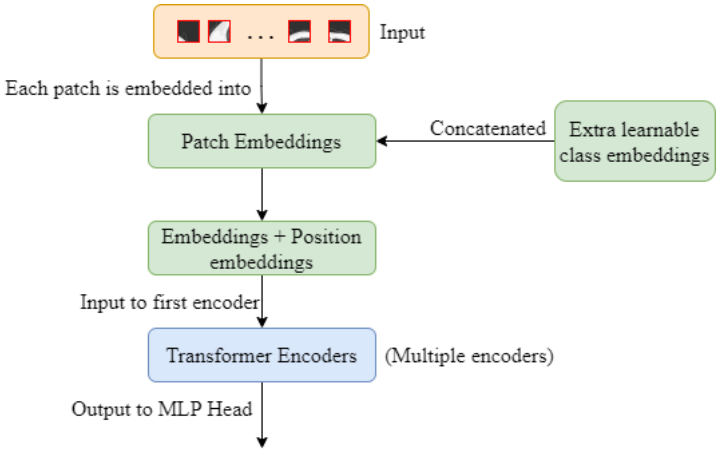

Figure 2. The Vision Transformer

Finally, the MSA is defined by considering h attention heads, h self-attention functions applied to the input, where each head provides a sequence of $N \times d$ and these h sequences are rearranged into a $N \times dh$ sequence re-projected by a linear layer into $N \times D$.

A ViT is composed by multiple trainable Encoder blocks (see Figure 3), where the first one receives as inputs the embedded image patches, and its outputs is fed into the next Encoder block and so on, until the last block's output is fed into a MLP head, which then classifies an image. Each one of these Encoder blocks is composed of two Batch Normalization layers and two residual layers, working together with the MSA and a MLP, as depicted on Figure 3.

The Vision Transformer (or any Transformer architecture) can take any kind of Encoder on its internal architecture. This flexibility made possible for researchers to come up with more efficient Encoder blocks, and therefore speeding up and occupying less memory. While the classic Encoder, used on the original Transformer paper [2] is feasible to be used with the ViT, we can still make it more efficient, and for that, we also experimented ViT with the Performer [4] encoder, an Encoder that tries to approximate the original one while maintaining linear time and space complexity. It consists of a novel approach called Fast Attention Via positive Orthogonal Random features(FAVOR+), which approximates softmax attention-kernels, allowing us to train our model faster with the same amount of parameters we used with a regular Transformer encoder.

For our last and most successful experiment, we used a pre-trained variation of the ViT, the Data-efficient image Transformer, or DeIT [11]. It aims to solve the large pre-training dataset requirement that the base ViT architecture has to deal with, due to its lack of inductive bias for 2D structures(images) by using the same ViT but using trained CNN architectures so that the ViT can inherit its knowledge and thus understand images a bit better, requiring less pre-train data because the Transformer network already has an inductive bias injected by the teacher CNN architecture.

We train Vision Transformer and Performer without transfer learning. For DeIT, we used the pre-trained weights from the original paper. For every one of these variations, we train with batches of 32 images per batch, resize the images to the required

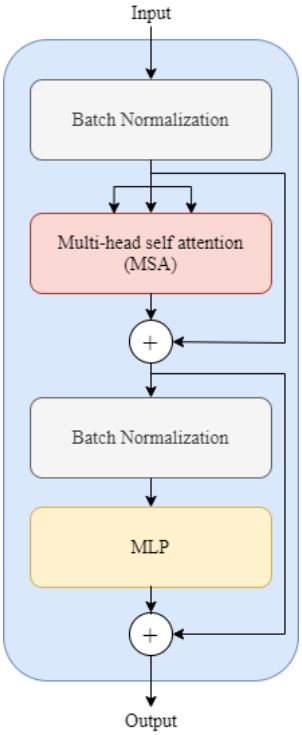

Figure 3. An encoder block

Table 2. Models Hyperparameteres

Hyperparameter	ViT w/ Transformer Encoder	ViT w/ Performer Encoder	DeIT Base
Learning Rate	6e-5	3e-5	8e-4
Dropout	0.5	0.5	0.5
Weight Decay	0	1e-4	1e-4
ViT's Patch Size	16	32	16
Number of Epochs	60	60	80

input size for each model (512x512 for Classical encoder ViT and Performer encoder ViT and 384x384 for DeIT) and apply some image augmentation techniques like randomly applying vertical and horizontal flips, random rotation and brightness variation. We also used an ADAM Optimizer and a linear learning rate scheduler for each model. For every epoch, we take the output and assess performance on the validation set, using the metrics described on Section 3, and at the end of training, we use the test set for the final evaluation. The main hyperparameters we used are described for each one of the models on Table 2.

3. Results

This sections focuses on the discussion of the results of our method, comparing its performance with other papers and then we discuss our achievements.

The Transformer block's output is fed into a simple MLP with one hidden layer, and is then translated into probabilities through a Softmax and then predicted into one of

Table 3. ViT with Classic Transformer Encoder(Accuracy: 0.91)

Class	Precision	Recall	F1-Score
Normal	0.95	0.98	0.97
Other Pneumonia	0.91	0.82	0.86
Covid-19 Pneumonia	0.81	0.88	0.84
Macro Average	0.89	0.89	0.89
Weighted Average	0.91	0.91	0.90

Table 4. ViT with Efficient Performer Encoder(Accuracy: 0.91)

Class	Precision	Recall	F1-Score
Normal	0.96	0.99	0.98
Other Pneumonia	0.84	0.92	0.88
Covid-19 Pneumonia	0.92	0.72	0.81
Macro Average	0.91	0.88	0.89
Weighted Average	0.91	0.91	0.91

the three classes. At the end of each training epoch, we assess the training performance on a validation set, and at the end of training, we evaluate our network on the test set.

For measuring our models performances, we used classic metrics for classification problems: accuracy, recall, precision and f1-score measured for each class.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \tag{2}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{4}$$

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

To address class imbalance, for recall, precision and f1-score, we also calculate its macro average, the unweighted mean for each one of the labels, and the weighted average, where it is calculated the weighted average.

Our Results on the test set, for 21191 CT images are detailed on Table 3,4, and 6, where we present precision and recall by class, macro and weighted average for each one of the tested model: ViT with Classic Transformer Encoder, ViT with Efficient Performer Encoder and Data efficient Vision Transformer, respectively.

We also compare with other works, with respect to accuracy. This comparison is shown on table 6. DenseNet201 based deep transfer learning only classify between COVID/NON-COVID, while the rest of the networks on the table classify between Healthy, Pneumonia, and pneumonia caused by COVID-19.

We obtained promising results, with the best one being from the pre-trained DeIT. Both ViT networks performed very well. With necessary pre-training, they are likely to

Table 5. Data efficient Vision Transformer(Accuracy: 0.96)

Class	Precision	Recall	F1-Score
Normal	0.98	0.99	0.99
Other Pneumonia	0.96	0.93	0.95
Covid-19 Pneumonia	0.92	0.94	0.93
Macro Average	0.96	0.96	0.96
Weighted Average	0.96	0.96	0.96

Table 6. Overall comparison with other works

Network	Accuracy
DenseNet201 based deep transfer learning[14]	0.96
Q-Deformed Entropy[13]	0.99
CovidNET-CT[5]	0.99
ViT with Classic Transformer Encoder	**0.91**
ViT with Efficient Performer Encoder	**0.91**
Data efficient Vision Transformer	**0.96**

surpass DeIT's performance, as shown on ImageNet's top-performing networks [1].

As observed in the overall results (Table 6), even though our accuracy is smaller than every other work, we obtained competitive results for a novel neural network that was initially meant for NLP problems, showing that the ViT modification can be used for general purpose computer vision problems with none to minimal modification of the original network, while other works were tailored specifically for the problem of classifying COVID-19 images.

Our metrics for COVID-19 classification were a little worse than the others, with healthy image classifications showing almost 100% f1-score, which is expected, due to class imbalance.

We can also observe that the difference between a ViT with a classical transformer encoder and one with an efficient performer encoder is minimal, meaning that the Performer encoder does approximate well the Transformer encoder, as stated on [4], while requiring less computing resources. The work we've done can be easily extended and fine-tuned for new kinds of viruses that may appear, resulting in a fast and precise low-sensitivity diagnosis tool. As the computation power available was limited, we didn't do the experiments using the original, pre-trained, top-performing Vision Transformer, ViT-Huge(ViT-H). Instead, we obtained a custom implementation of the network and tuned its hyperparameters to extract its potential.

4. Conclusion

This work proposed the use of the transformer architecture for medical image classification. Even though COVID-19 diagnosis can be made with more traditional methods, computed tomography can still be a significant financial or structural obstacle for some parts of the society, but this type of equipment is getting more and more accessible for

[1] Benchmark on ImageNet at https://paperswithcode.com/sota/image-classification-on-imagenet

people. To extract the most of this technology, we can use it to help radiologists filter higher-risk patients who show signs of pneumonia caused by COVID-19 or some other dangerous virus.

We compared important metrics between different Deep Learning approaches, exploring the novel transformer and comparing it with classical CNNs. We demonstrate that modest, pre-trained models like DeIT can achieve competitive results, and the original Vision Transformer, even without pre-training and no tailoring for the task at hand can also achieve some good results. We hope our paper motivates other researchers to explore the Transformer Architecture for images and medical diagnosis.

For future work, we propose the use of pre-trained weights from Vision Transformer, but with a Performer Encoder, making the training faster and less computationally expensive.

5. Acknowledgements

The authors acknowledge the Brazilian institutions: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) for the financial support.

References

[1] Alsharif and A. Qurashi, Effectiveness of COVID-19 diagnosis and management tools: A review, Radiography, 10.1016/j.radi.2020.09.010

[2] Vaswani, Ashish, et al.: Attention is all you need. Advances in neural information processing systems. 2017. https://arxiv.org/abs/1706.03762

[3] Dosovitskiy, Alexey, et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020). https://arxiv.org/abs/2010.11929

[4] Choromanski, Krzysztof, et al.: Rethinking attention with performers. arXiv preprint arXiv:2009.14794 (2020). https://arxiv.org/abs/2009.14794

[5] Gunraj, Hayden, Linda Wang, and Alexander Wong.: Covidnet-ct: A tailored deep convolutional neural network design for detection of covid-19 cases from chest ct images. Frontiers in Medicine 7 (2020). https://arxiv.org/abs/2009.05383

[6] Afshar, Parnian, et al.: Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. https://arxiv.org/abs/2004.02696

[7] Abbas, Asmaa, Mohammed M. Abdelsamea, and Mohamed Medhat Gaber.: Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. (2020). https://arxiv.org/abs/2003.13815

[8] He, Kaiming, et al.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. https://arxiv.org/abs/1512.03385

[9] Tan, Mingxing, and Quoc V. Le.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019). https://arxiv.org/abs/1905.11946

[10] Silva, Pedro, et al.: COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. Informatics in Medicine Unlocked 20 (2020): 100427. https://www.sciencedirect.com/science/article/pii/S2352914820305773

[11] Touvron, Hugo, et al.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020). https://arxiv.org/abs/2012.12877

[12] Gunraj, Hayden, et al.: COVIDNet-CT: A COVID-19 CT Dataset. https://github.com/haydengunraj/COVIDNet-CT

[13] Hasan, A.M.; AL-Jawad, M.M.; Jalab, H.A.; Shaiba, H.; Ibrahim, R.W.; AL-Shamasneh A.R.: Classification of Covid-19 Coronavirus, Pneumonia and Healthy Lungs in CT Scans Using Q-Deformed Entropy and Deep Learning Features. Entropy 2020, 22, 517. https://doi.org/10.3390/e22050517

[14] Aayush Jaiswal, Neha Gianchandani, Dilbag Singh, Vijay Kumar Manjit Kaur (2020) Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1788642