JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DIAGNOSIS OF COVID-19 USING X-RAY IMAGES AND CONVOLUTIONAL NEURAL NETWORK AND VISION TRANSFORMER

Comparitive Analysis of CNN and ViT for Covid-19 diagnosis

[1]Zulqarnain Ahmed, [2]Dr. M A Anusuya
[1]M.Tech Computer Engineering, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]JSS Science and Technology University, Mysuru, India

***Abstract*:** With the onset of COVID-19 pandemic that affected millions of lives across the globe and was declared as a medical health emergency by the World Health Organization (WHO), the main challenge that was present was to diagnose the disease correctly so that its effects could be mitigated by suitable medicines. In this paper we take a look at a relatively new method of image classification i.e. Vision Transformer (ViT) [1] introduced by Google AI Labs [2] in 2017 and compare it with the prevalent method of Convolutional Neural Networks (CNNs) in medical imagery. Due to the non-availability of sufficient-size and good-quality chest X-ray image dataset, an effective and accurate classification was a challenge. To deal with these complexities such as the availability of a very-small-sized and imbalanced dataset with image-quality issues, the dataset has been preprocessed in different phases using different techniques to achieve an effective training dataset for the CNN and vision transformer models to attain its best performance. The preprocessing stages of the datasets performed in this study include dataset balancing, and data augmentation.

In this work we have used the confusion matrix metrics such as accuracy, precision, recall, specificity, and F1-score to compare the results obtained by the two models. The experimental results have shown the overall accuracy as high as 86.5% for CNN and 96.66% for ViT model which demonstrates the good capability of the proposed ViT model in the current application domain and how vision transformers are impacted by dataset size and training resources.

Keywords— Convolutional neural networks, vision transformers Keras, Tensorflow, COVID19, pandemic, deep learning, confusion matrix

I. INTRODUCTION

The COVID-19 pandemic took its toll on the existing health systems all over the world. Numerous ways of detecting and diagnosing the disease were proposed with the most prevalent of them being the reverse transcription polymerase chain reaction popularly known as RT-PCR tests. The turnaround times for these methods varied between a couple of hours to a day. Many researches also proposed the use of X-ray or CT scan images of the person's chest to diagnose the disease. In the medical imaging applications CNN remains one of the top go to methods for various applications such as segmentation, image detection, classification, etc. In our study we mainly focus on a relatively new method called Transformer Neural Network that was proposed by Google in 2017. Transformers are based on the attention mechanism and also account for the long range dependencies between words in a sentence or a pixel or group of pixels in an image. Transformers have become very popular in natural language processing (NLP) applications such as text translations and summary. IT has replaced RNNs and LSTMs in the NLP domain. Use of transformers in image applications has been explained in [3]. We use this approach to train our model to detect COVID-19 using the supplied X-ray images and compare it with CNN model.

II. METHODOLOGY

A. Datasets used

- A public open-source dataset made available by Dr. Joseph Paul Cohen [4] containing numerous samples of lung diseases and their corresponding CXR images. The training set consists of 112 images each of COVID19 samples and normal healthy patients. The validation or testing set consists of 30 images each of normal and infected patients.

- "COVID-19 Radiography Database" [5] available in Kaggle was also used. This dataset is broader and contains more images than that of the previous dataset. The training set consists of 3458 images each for covid and normal samples. And 50 images each for testing and validation set each for normal and covid samples.

B. Data Preprocessing

Due to unavailability of large size dataset of chest X-rays particularly for COVID patients, we perform data augmentation on the existing dataset. Random flip, rotation, resizing and random zoom are applied on the images to generate more images that can be used to train and validate the model.

C. Feeding the image to the transformer

The ViT operates on patches of an image. Firstly we take the image and use the extract_patches function of tensorflow. In our approach we have divided each image of size 72 x 72 into patches of 6 x 6. This means each image will be split into 144 patches and fed to the transformer. These patches that are created are two dimensional and must be flattened to be fed linearly to the transformer.

D. Adding positional embeddings

For the transformer to map each patch to its position in the original image, the patches are embedded with a positional encoding starting from 1 to 144. These encodings are used by the encoder module of the transformer to calculate the attention scores of each patch with respect to all other patches of the image.

E. Training the model

We train the model in a fully supervised way by feeding labelled images to the transformer. The images are labelled as COVID or Normal. The transformer passes the input image to multiple encoders and the attention score obtained with the encoder is then compared to the target image fed to the decoder. We use a multi- head attention layer to calculate the attention scores of each patch with all the other patches seen so far by the neural network. In our approach we have used 4 heads in this module. This means that at a given point the patch can be operated parallely by 4 heads of the module therefore speeding up the training.

The final softmax layer in the MLP produces the final classification by converting the attention scores to probabilities ranging from 0 to 1.

F. Evaluating the model

The model is then evaluated by passing the validation or testing set of the data that the model has not seen during training. We then construct a confusion matrix that shows how many images were correctly classified and how many were incorrectly classified. The same approach is taken for evaluating a CNN model that has been trained with the same dataset and compute power.

III. RESULTS AND EVALUATION

To understand the effect of dataset size and resources on both CNN and vision transformer, we have divided our study into five cases. The details of each case is given below along with their results and confusion matrix metrics.

A. Case 1: CNN model with small dataset

In our first case we use a CNN model made of an input layer and an output layer and two hidden layers. We have used the max pooling between layers to aggregate the result of convolution operation between each layer. The dataset used for training consists of 224 images for training set and 60 images for the validation and testing set which have been Split into 112 images each for training the model for COVID and normal patients and 30 images each for COVID and normal patients.

The number of epochs during training that was found to give stable and high accuracy was around 10 for the CNN model. Hence the number of epochs chosen was 10.

The confusion matrix for case 1 is shown in figure 1. The confusion matrix metrics of all cases have been summarized in the table 1.

```
sns.heatmap(cn,cmap="plasma",annot=True) #0: Covid ; 1: Normal
<AxesSubplot:>
```

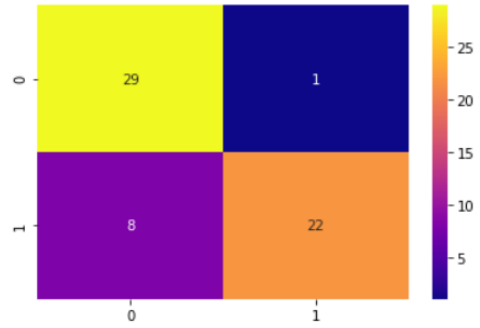

Figure 1: Confusion matrix for case 1

B. Case 2: ViT with small dataset

The data size taken is 224 images for training set and 60 images for the validation set. The data is further split into 112 images each for training the model for COVID and normal patients and 30 images each for COVID and normal patients. The epoch used was 100 for ViT to achieve better results. The confusion matrix for case 2 is shown in figure 2.

```
sns.heatmap(cn,cmap="plasma",annot=True) #0: Covid ; 1: Normal
<AxesSubplot:>
```

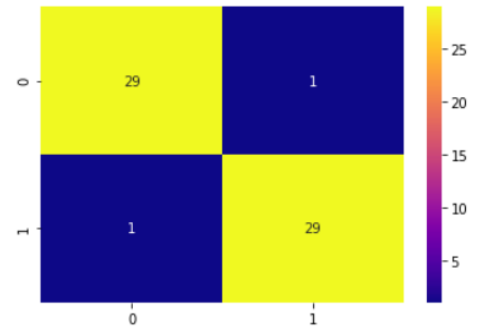

Figure 2: Confusion matrix for case 2

C. Case 3: CNN with large dataset

The data size taken is 6916 images for training set and 100 images for the validation set. The data is further split into 3458 images each for training the model for COVID and normal patients and 50 images each for COVID and normal patients. The number of epochs used was 10 for CNN to achieve better results.

```
sns.heatmap(cn,cmap="plasma",annot=True) #0: Covid ; 1: Norma
<AxesSubplot:>
```

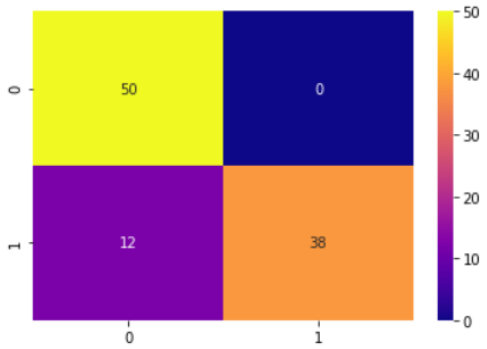

Figure 3: Confusion matrix for case 3

D. Case 4: ViT with large dataset keeping the epochs same as earlier

The data size taken is 6916 images for training set and 100 images for the validation set. The data is further split into 3458 images each for training the model for COVID and normal patients and 50 images each for COVID and normal patients. The epoch used was 100 for ViT to compare it to the earlier case.

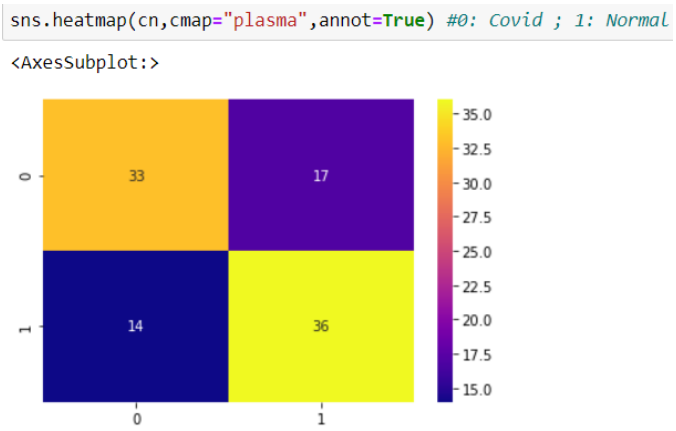

Figure 4: Confusion matrix for case 4

E. Case 5: ViT with large dataset and increased epochs

The data size taken is 6916 images for training set and 100 images for the validation set. The data is further split into 3458 images each for training the model for COVID and normal patients and 50 images each for COVID and normal patients. The epoch used was 200 for ViT to achieve better results.

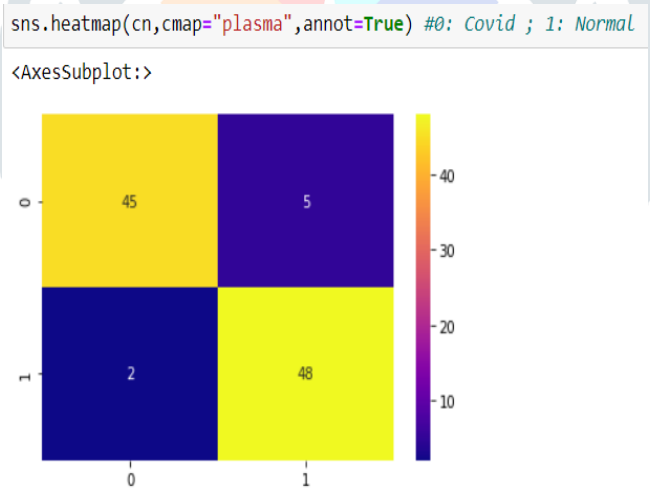

Figure 5: Confusion matrix for case 5

Table 1: Summary of confusion matrix metrics obtained across all cases

Case	Algorithm	Epochs	Dataset		Accuracy	Precision	Recall	Specificity	F-Score
			Training	Testing					
Case 1	CNN	10	224	60	85%	78.37%	96.66%	73.33%	86.56%
Case 2	ViT	100	224	60	96.66%	96.66%	96.66%	96.66%	96.66%
Case 3	CNN	10	6916	100	88%	80.64%	100%	76%	89.28%
Case 4	ViT	100	6916	100	69%	70.21%	66%	72%	68.04%
Case 5	ViT	200	6916	100	93%	95.74%	90%	96%	92.78%

F. Summary of the result

- Through CNN, even large datasets converge faster as we see lesser number of epochs are needed to achieve a good accuracy and F-Score percentage. This fact is supported by case 1 and case 3 where we have a smaller dataset and larger dataset respectively and the model worked with similar results with not much change in both the accuracy and F1-Score. On the other hand, in our vision transformer model we see that it performs well in case 2 with a smaller dataset and 100 epochs. When the dataset size is increased in case 4 the accuracy drops to 69% which is not desirable in a medical diagnostic application. In case 5 when the number of epochs is increased for the ViT model with a bigger dataset the accuracy reaches competitive levels. Thus the ViT model needs more time to converge with larger datasets.

- The variance found in ViT models is more because it responds to change in dataset more than the CNN model. In cases 1 and 3 for CNN not much change is seen but that is not the case for ViT model as seen in cases 2 and 4.

- Comparing both models, we see that ViT achieves a better accuracy and F1-Score compared to CNN.

IV. Conclusion and Future Work

Detecting the presence of SARS-COVID virus from chest X-ray images at an early stage is not an easy task as the signs are not clearly visible thus leading to severe complexities such as inflamed lungs, nausea, loss of olfactory senses and severe pneumonia and cold. This project aimed at improving the performance of machine learning and deep learning classifiers in the detection of COVID19 virus from chest X-ray images. The classification was improved by considering various aspects which affect the performance; such as the distribution of data among classes in the dataset and the quality of images being used to train the classifiers. With larger dataset we saw how it affects the vision transformer and CNN model.

To counter the shortage of medical image datasets for this application we also used data augmentation to increase the dataset size.

Although our model classifies COVID-19 chest X-rays with leading accuracy, there is still room for improvement. The performance of the model in radiography is significantly impacted by noise. The performance of our model can be significantly enhanced by using noise reduction methods based on Generative Adversarial Networks (GAN) [6] on the dataset. The final performance can be improved by using an ensemble learning-based technique. Higher performance metrics can definitely be attained by using a large version of ViT [7] and a larger dataset. Additionally, this study can be expanded to segment COVID-19 chest X-rays and CT images to provide radiologists with even more information.

V. Acknowledgment

Through this column, I would like to take this opportunity to thank all those who have inspired and guided me to successfully complete my work on this research. I thank God for His blessings upon me and making this work easier for me through his immense mercy. I am grateful to the principal **Dr. S B Kivade** for letting me be a part of this prestigious institution and letting us explore our abilities to the fullest. I also express my sincere gratitude towards our HOD, **Dr. Srinath S** for giving us the platform to enhance our knowledge and research and present our work as a final year project. I would like to express my sincere gratitude to my guide, **Dr. Anusuya M A** for her patience, intense knowledge, motivation and guidance throughout our work. I would also like to thank my family members for their continuous support and cooperation.

References

[1] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, doi: 10.48550/arxiv.2010.11929.

[2] A. Vaswani *et al.*, "Attention Is All You Need," *Adv. Neural Inf. Process. Syst.*, vol. 2017–December, pp. 5999–6009, Jun. 2017, doi: 10.48550/arxiv.1706.03762.

[3] M. E. H. Chowdhury *et al.*, "Can AI help in screening Viral and COVID-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, Mar. 2020, doi: 10.1109/ACCESS.2020.3010287.

[4] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 Image Data Collection," Mar. 2020, doi: 10.48550/arxiv.2003.11597.

[5] "COVID-19 Radiography Database | Kaggle." https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (accessed Jul. 05, 2022).

[6] Y. Sun, X. Liu, P. Cong, L. Li, and Z. Zhao, "Digital radiography image denoising using a generative adversarial network," *J. Xray. Sci. Technol.*, vol. 26, no. 4, pp. 523–534, 2018, doi: 10.3233/XST-17356.

[7] C. Park, Y. Jeong, M. Cho, and J. Park, "Fast Point Transformer," Dec. 2021, Accessed: Jul. 05, 2022. [Online]. Available: http://arxiv.org/abs/2112.04702