© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Investigation of ConViT on COVID-19 Lung Image Classification and the Effects of Image Resolution and Number of Attention Heads

Pun Liang Thon[1], Joel C. M. Than[1*], Norliza M. Noor[2], Jun Han[3], Patrick Then[1]

[1]Faculty of Engineering, Computing and Science,
Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, Kuching, 93350, MALAYSIA

[2]Razak Faculty of Technology and Informatics,
Universiti Teknologi Malaysia, Jalan Iman, Skudai, 81310, MALAYSIA

[3]Department of Computing Technologies,
Swinburne University of Technology Hawthorn Campus, John Street, Hawthorn, Victoria 3122, AUSTRALIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2023.15.03.005
Received 30 October 2022; Accepted 29 December 2022; Available online 31 July 2023

Abstract: COVID-19 has been one of the popular foci in the research community since its first outbreak in China, 2019. Radiological patterns such as ground glass opacity (GGO) and consolidations are often found in CT scan images of moderate to severe COVID-19 patients. Therefore, a deep learning model can be trained to distinguish COVID-19 patients using their CT scan images. Convolutional Neural Networks (CNNs) has been a popular choice for this type of classification task. Another potential method is the use of vision transformer with convolution, resulting in Convolutional Vision Transformer (ConViT), to possibly produce on par performance using less computational resources. In this study, ConViT is applied to diagnose COVID-19 cases from lung CT scan images. Particularly, we investigated the relationship of the input image pixel resolutions and the number of attention heads used in ConViT and their effects on the model's performance. Specifically, we used 512x512, 224x224 and 128x128 pixels resolution to train the model with 4 (tiny), 9 (small) and 16 (base) number of attention heads used. An open access dataset consisting of 2282 COVID-19 CT images and 9776 Normal CT images from Iran is used in this study. By using 128x128 image pixels resolution, training using 16 attention heads, the ConViT model has achieved an accuracy of 98.01%, sensitivity of 90.83%, specificity of 99.69%, positive predictive value (PPV) of 95.58%, negative predictive value (NPV) of 97.89% and F1-score of 94.55%. The model has also achieved improved performance over other recent studies that used the same dataset. In conclusion, this study has shown that the ConViT model can play a meaningful role to complement RT-PCR test on COVID-19 close contacts and patients.

Keywords: COVID-19, convolutional vision transformer, deep learning, disease classification

1. Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic has been an ongoing battle in many countries since its outbreak in December 2019 in Wuhan, China. To date, the COVID-19 recorded confirmed cases has surpassed 400 million and has caused over 5.8 million death worldwide [1]. Since then, the virus has evolved into different variants [2], including Delta and Omicron, contributing to the major cases of COVID-19 in Malaysia. Besides vaccination, one key towards combating this pandemic is timely, effective and accurate diagnosis of the disease in persons under investigation and close contacts. To diagnose COVID-19, the rapid antigen test (RAT) is the most popular test due to its high

*Corresponding author: jcmthan@swinburne.edu.my

accessibility, which is available in many pharmacies and convenience stores, affordable and fast result time of about 10 to 30 minutes. However, the RAT comes with very low sensitivity, which may be due to the low viral loads of the test sample [3]. Another test available is the reverse transcription polymerase chain reaction (RT-PCR) test, which is the golden standard for diagnosing COVID-19 due to its high sensitivity and specificity [4], [5]. However, the test is not easily accessible and has a high turnaround time.

The use of radiological imaging such as chest computed tomography (CT) and chest X-rays (CXR) in diagnosing COVID-19 patients is another viable option for diagnose and follow-up treatment of COVID-19 patients. This is due to the radiological presentation and patterns found in COVID-19 infected lung image, such as broncho vascular thickening (in lesions), ground glass opacity and peripheral distribution [6], [7] which is not presented in normal or healthy lung image. These differences have allowed the use of deep learning to diagnose COVID-19 using chest imaging.

The use of Convolutional Neural Networks (CNNs) plays a significant role in image classification tasks, including disease diagnosis using medical images. Many recent works have been done on using CNNs to detect COVID-19 positive cases, including but not limited to the utilization of CT and CXR images. Many researchers favour CNNs as it has hard inductive biases which enable sample-efficient learning. This is meaningful because medical imaging data are scarce, and many are not open access due to privacy consideration and is not cost-efficient to build. However, CNNs often require costly computational resources to train and are said to have a lower performance ceiling [8].

This leads to introducing the Vision Transformer (ViT) [9] which purely uses attention for computer vision tasks instead of using convolution. Thus, as compared to CNNs, ViT demands fewer computational resources to train. However, without convolution, ViT lacks hard inductive biases. In order for ViT to achieve performance on par if not better than CNNs, it is required to train on large external datasets [8]. Specifically, in medical imaging analysis problems, researchers face difficulties in getting huge datasets to train on ViT to achieve as good results as by CNNs as medical data are limited and expensive.

The introduction of Convolutional Vision Transformer (ConViT) successfully solve this problem by introducing soft inductive bias in ViT [8]. The authors introduced *gated positional self-attention* (GPSA) layers, that is able to mimic the locality of convolutional layers, and then allow each attention head to escape the locality. ConViT has proven to combine the strength of CNNs and ViT, which are inductive bias and cheap training cost, and avoid their limitations. This study investigates the effects of image pixels resolution used to train the model and the number of attention heads used in the model on COVID-19 CT scan image data. The original work assumed that using 224x224 pixels resolution and 16 attention heads gives the best performance.

While there are many studies utilised ViT and ConViT in classification tasks, to the best of our knowledge, the effects of different image pixel resolutions and the number of attention heads are not clearly investigated in related studies. In this study, ConViT is applied to diagnose COVID-19 cases from lung CT scan images. Particularly, we investigated the relationship of the input image pixel resolutions and the number of attention heads used in ConViT and their effects on the model's performance in COVID-19 diagnosis task. The best model in this experiment is compared with other state-of-the-art methods that uses the same dataset and has shown improved performance.

2. Related Work

Many works have proved that CT scan images can be used in deep learning to diagnose the COVID-19 disease. Among them, the use of CNNs is a popular choice of application. Rahimzadeh et al. introduced a new feature pyramid network to the ResNet50V2 model to allow investigation on different image resolutions and avoid losing tiny objects, which increases the classification performance significantly [10]. Training on 15589 COVID-19 CT images and 48260 normal CT images, their proposed model has achieved an overall accuracy of 98.49%. Similiarly, Matsuyama adopted the use of wavelet coefficients in a ResNet-50 based model to differentiate COVID-19 cases and non-COVID-19 cases from chest CT scan images [11]. Using a dataset with only 720 chest CT scan images (345 COVID-19 and 375 non-COVID-19), the proposed method has achieved an accuracy of 92.2%. Similarly, He et al. introduced the Self-Trans approach, which integrates contrastive self-supervised learning with transfer learning [12]. Trained on CRNet proposed by the authors on 249 COVID-19 positive CT scan images and 397 negative CT scan images with the Self-Trans approach, the authors have achieved an overall accuracy of 86%. It was observed that using smaller datasets, CNNs could still produce promising performance.

Besides CNNs, the utilisation of ViT for COVID-19 diagnosis also gains its popularity since its release. Mondal et al. utilised ViT for COVID-19 screening using both chest CT and CXR images [13]. To address data scarcity that will affect the ceiling performance of ViT due to the lack of inductive bias, they employed a multi-stage transfer learning technique. Besides, the authors used the Gradient Attention Rollout algorithm [14] to make the model explainable. Trained on 143778 total chest CT images (35996 Normal, 25496 Pneumonia and 82286 COVID-19), the model has achieved an overall accuracy of 98.1%; trained on 5911 total CXT images (1079 Normal, 3106 Pneumonia and 1726 COVID-19), the model has achieved an overall accuracy of 96%. Similiarly, Park et al. proposed a novel ViT that uses low-level CXR features corpus obtained from a backbone network that extracts common CXR findings [15]. Trained and tested on multiple external datasets, their model has achieved a peak overall accuracy of 93.2% in three-classes (Normal, Others and COVID-19) classification tasks. Gao et al. compared the performance between ViT and DenseNet in

classifying COVID-19 and non-COVID-19 based on chest CT images [16]. The performance of both models is compared after being trained on 1552 CT 3D subjects (687 COVID-19, 865 non-COVID-19). ViT has achieved a better F1-score of 0.76 than DenseNet of 0.72. Meanwhile, Mehboob et al. has developed a self-attention transformer-based approach for COVID-19 diagnosis [17]. The proposed approach is compared with CNN-based and Ensemble-based classifiers using two different CT scan datasets. Their proposed method is proven more effective in detecting COVID-19 with a peak accuracy of 99.7% on multi class HUST-19 CT scan dataset (4001 COVID-19, 9979, non-COVID-19, 5705 non-informative).

In this work, we investigated the use of ConViT, combination of strengths of CNNs and ViT, in diagnosing COVID-19 patients from lung CT images, using different image pixels resolution and number of attention heads.

Table 1 - Summary of related studies on COVID-19 diagnosis using CNNs and ViT on CT scan images

Year	Model	Classes	Accuracy
2021	ResNet50V2 + Feature Pyramid Network [10]	COVID-19, Normal	98.49%
2020	ResNet-50 + Wavelet Coefficients [11]	COVID-19, non-COVID-19	92.2%
2020	CRNet + Self Trans [12]	COVID-19, non-COVID-19	86%
2022	xViTCOS with multi-stage transfer learning [13]	COVID-19, Normal, Pneumonia	98.1%
2021	ViT [15]	COVID-19, Normal, Others	93.2%
2021	ViT [16]	COVID-19, non-COVID-19	76.6%
2022	Vision Self-Attention Transformer [17]	COVID-19, non-COVID-19, non-informative	99.7%

3. Methodology

3.1 Dataset

To demonstrate the use of ConViT in COVID-19 diagnosis under different configurations and different dataset pixel sizes, we used an large open source COVID-19 dataset, COVID-CTset [10]. The dataset contains 15589 CT scan images from 95 COVID-19 patients and 48260 CT scan images from 282 normal persons, all gathered from the Negin medical center in Sari, Iran. The lung HRCT images were obtained from the patients using a SOMATOM Scope model and syngo CT VC30-easyIQ software version. They were readied in TIFF format with 512x512 pixels resolution and holds the same 16-bit grayscale data.

Since the dataset is very large, the dataset authors also prepared a smaller dataset that is part of the main dataset which is sufficient for training and testing deep networks. Therefore, in this study, the subset of 12059 (2282 COVID-19, 9776 Normal) of 63849 total images were used. Examples of the COVID-19 CT scan image are shown in figure 1 and Normal CT scan image are shown in figure 2.

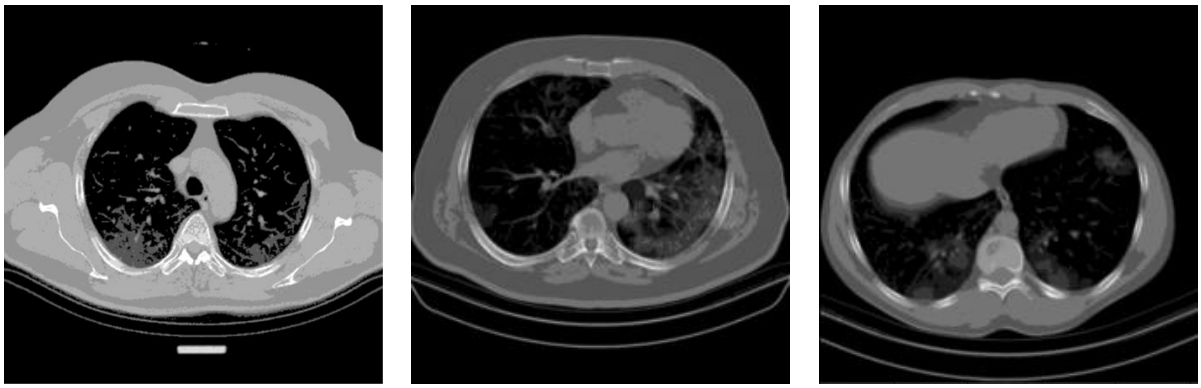


Fig. 1 - Examples of COVID-19 CT images used in this studies

Fig. 2 - Examples of Normal CT images used in this studies

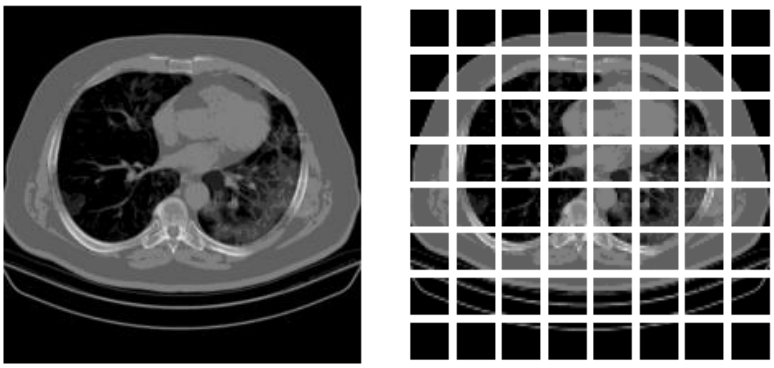
The dataset is split into 80% train set, 10% validation set and 10% test set. The important statistics of the dataset can be found in Table 2. Next, the dataset is resized to 224x224 pixels resolution and 128x128 pixels resolution for our experiment.

Table 2 - Summarised description of the dataset

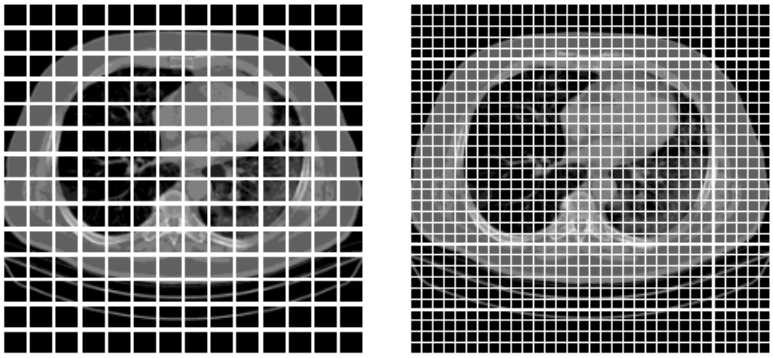
Split	COVID-19	Normal	Total
Train	1825	7820	9645
Validation	228	977	1205
Test	229	979	1208

3.2 Patch Segregation

Each images in the dataset is divided into patches, each patch contains 16x16 pixels. The higher the image pixels resolution, the more patches will be produced during division. Figure 3 shows images with different pixels resolution segregated by 16x16 patch size. Images with 128x128 pixels resolution will produce 64 patches per image when segregated by 16x16 patch size (Figure 3(b)); Images with 224x224 pixels resolution will produce 196 patches per image (Figure 3(c)); Images with 512x512 pixels resolution will produce 1024 patches per image (Figure 3(d)). In this study, the best pixels resolution of the dataset for training and testing the ConViT model is investigated. The size of the patch and the areas of lung tissue covered in a patch contains an amount of global and local information that plays a factor in the model performance later.



(a) (b)



(c) (d)

Fig. 3 - COVID-19 CT scan images in different pixels resolution segregated by 16x16 patch size (a) Original image; (b) 128x128 pixels resolution; (c) 224x224 pixels resolution; (d) 512x512 pixels resolution

3.3 Convolutional Vision Transformer (ConViT)

ConViT [8] is built based on ViT [9] and integrated the strengths of CNNs. ViT uses only transformers, which makes the network architecture less complicated, requires less GPU resources and faster to train as compared to CNNs. First, the input image is segregated into smaller fixed patches, and linearly embed each of the patches. Then, positional embeddings are appended to the vector to preserve positional information of these patches. The product is then feed to a standard transformer encoder, which consists of multiheaded self-attention layers and multilayer perceptron (MLP) units [18]. A normalisation layer, or Layernorm (LN) is applied before and after every multiheaded unit [19]. Finally, the MLP head will output the class label prediction of the input image using the product of the transformer encoder.

Figure 4 shows the overview of the ConViT architecture. Instead of using purely transformer, d'Ascoli et al. introduced the GPSA layers to introduce "soft" inductive biases to ViT, resulting in ConViT [8]. The GPSA layers are initialized to simulate the locality of convolutional layers, and then they can decide whether to escape locality or not. This is done by tuning a gating parameter in the layer which regulates the attention paid to position versus content information. This allows application of "soft" inductive biases instead of typical hard inductive biases used in CNNs.

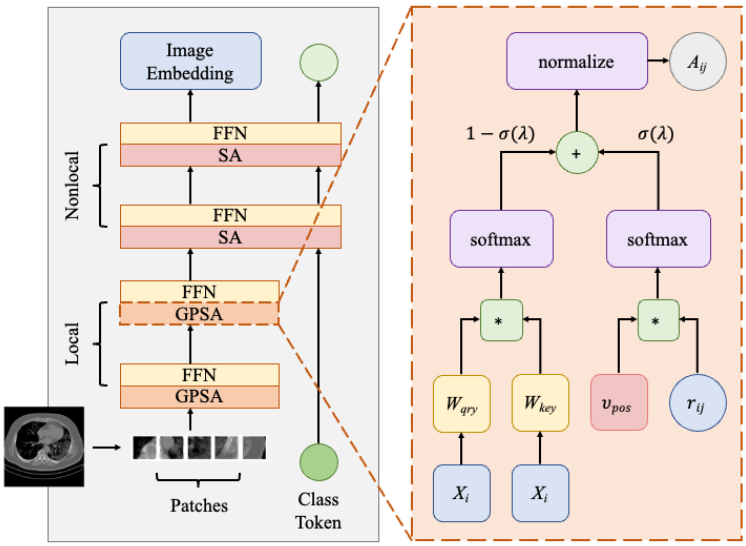


Fig. 4 - Overview of the ConViT architecture

In this experiment, similarly, the ConViT slices the input images by 16x16 patch size. The patches are projected linearly, resulting in a sequence of vectors. Then, positional embeddings are added to the vector. These patches are then propagated through 12 blocks in ConViT. Each of the first 10 blocks consists of GPSA layers followed by a 2-layer Feed-Forward Network (FFN) with GeLU activation. The final 2 blocks are consisting of SA layers followed by another 2-layer FFN. Unlike ViT, the class token of the input image patches is only appended after the last GPSA layer. This is because GPSA layers involve positional information, thus the class token approach will be redundant.

3.4 Other Experiment Configurations

The batch size used is 32. Cross-entropy loss function is employed. Adam optimiser is used with a learning rate of 0.00003. StepLR is also employed throughout the experiments. It updates the learning rate after every step size by gamma, where in this study, the learning rate is updated every epoch by multiplying the current learning rate to gamma, where gamma is 0.7.

3.5 Performance Evaluation

We compute and report accuracy (1), sensitivity (2), specificity (3), positive predictive value (PPV) (4), negative predictive value (NPV) (5) and F1-score (6) to evaluate the performance of the ConViT model under different number of attention heads used and different CT scan image resolutions. The positive class represents COVID-19 lung CT scan images whereas the negative class represents normal lung CT scan images.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \times 100\% \quad (3)$$

$$\text{PPV} = \frac{TP}{TP + FP} \times 100\% \quad (4)$$

$$\text{NPV} = \frac{TN}{TN + FN} \times 100\% \quad (5)$$

$$\text{F1-Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \times 100\% \quad (6)$$

These performance metrics are calculated from True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) after the model is evaluated on the test dataset. Having several performance metrics helps to evaluate the performance of the classification model from different perspectives. Using the performance metrics mentioned, we also compared the best performing model in this experiment with other state-of-the-art methods that are using the same dataset [10], [20].

4. Result and Discussion

In this study, we present, analyse, and discuss the result of this study. Table 3 presents the performance of ConViT under different number of attention heads and different CT scan image resolutions used to train the model. As shown in the table, using 128x128 pixels resolution and 16 attention heads in ConViT yielded the highest accuracy (98.01%). This indicator is supported by the other performance metrics, with sensitivity of 90.83%, specificity of 99.69%, PPV of 98.58%, NPV of 97.89% and F1-score of 94.55%, which achieved the highest numbers among other configurations. In general, all models achieved higher specificity than sensitivity and higher PPV than NPV. This may be due to the imbalance number of COVID-19 CT images (1825 images) and Normal CT images (7820 images) used to train the model.

From Table 3, it is shown that the best model has achieved a high value of sensitivity at 90.83%, which implies that a small portion of COVID-19 CT scan images are incorrectly identified as Normal CT scan images. However, this small portion of FN could lead to the spread of the COVID-19 as the patient who genuinely has the disease might get rejected or delayed for treatment and quarantine. The best model also attains very high value of specificity of 99.69%, which indicates the extremely low number of FP. In this COVID-19 pandemic, this is important to avoid individuals who are healthy from unintentionally occupying medical resources from those who really needs them, i.e., COVID-19 patients. Besides, the model has achieved high value of PPV (98.58%) and NPV (97.89%) which indicates that the predictions made by the model are very likely to be correct. F1-score of 94.55% also indicates that the model has achieved the most balance between PPV (also known as precision), and sensitivity (also known as recall).

Table 3 - Performance of ConViT under different number of attention heads and different pixels resolutions

Resolution	Type (Number of Attention Heads)	Performance (%)					
		Accuracy	Sensitivity	Specificity	PPV	NPV	F1-Score
512x512	Tiny (4)	92.38	67.25	98.26	90.06	92.77	77.00
	Small (9)	93.38	74.24	97.85	89.01	94.20	80.95
	Base (16)	92.88	68.12	98.67	92.31	92.97	78.39
224x224	Tiny (4)	94.54	72.93	99.59	97.66	94.02	83.50
	Small (9)	94.70	86.46	96.63	85.71	96.83	86.09
	Base (16)	96.27	86.46	98.57	93.40	96.89	89.80
128x128	Tiny (4)	92.80	82.10	95.30	80.34	95.79	81.21
	Small (9)	97.60	90.39	99.28	96.73	97.79	93.45
	Base (16)	**98.01**	**90.83**	**99.69**	**98.58**	**97.89**	**94.55**

Overall, it is observed that with bigger image pixels resolution, the classification performance drops. This suggests the use of patches for ConViT to classify images is effective. It is proven in the original paper of ViT [9] that segregating images with pixels resolution of 224x224 by 16x16 patch size, resulting in a total of 196 patches per image, has achieved the best performance in classification problem. However, in this study, segregating images with pixels resolution of 128x128 by 16x16 patch size, resulting in a total of 64 patches per image, has outperformed pixels resolution of 224x224. This observation suggests that the dataset used in this study works best at 128x128 pixels resolution segregated by 16x16 patch size most likely due to the balance of global and local information in each patch. As shown in figure 3, increasing pixels resolution results in producing more patches, encouraging the loss of global information in each patch which could affect the classification performance. This observation is in line with the experiment done on pixels resolution of 512x512, the performance of the classifier is the worst compared to other settings as the images are segregated by 16x16 patch size, producing 1024 patches per image, thus results in more vital global information loss required for classification.

From Table 3, looking at experiments done on each pixels resolution individually with different number of attention heads used, it is observed that as the number of attention heads used increases, the classification performance rises for the 224x224 pixels resolution and 128x128 pixels resolution; but for 512x512 pixels resolution, the overall accuracy drops from 93.38% to 92.88%, where number of attentions heads used increase from 9 to 16. This is the only exception in the experiment, which may be due to low number of training epochs considering 512x512 pixels resolution requires longer training to achieve stable performance. Overall, the increasing numbers of attention heads allows the model to attend to information from different perspectives together [18], thus increasing the number of learnable parameters and allow more interaction between patches. With this, performance is likely to increase as the number of attention heads used increase.

In figure 5, 6 and 7 shows the training curves of ConViT-B (16 attention heads) on 128x128, 224x224, 512x512 pixels resolution data over 20 epochs. It is observed that as the training data pixels resolution increase, the train loss and validation loss becomes more unstable. Besides, it is observed that the train and validation loss curves do not have a good fit. As the dataset pixels resolution increases, the model takes more epochs to reach stability on train and validation accuracy and loss. In figure 7, it is observed that there is a drop in accuracy and rise in loss at epoch 11 which may be due to accidentally stopping of the model training. From all three figures, it is therefore concluded and suggested that as the dataset pixels resolution increases, the model should be train on more epochs until it has a good fit learning curves to ensure the peak performance of the model.

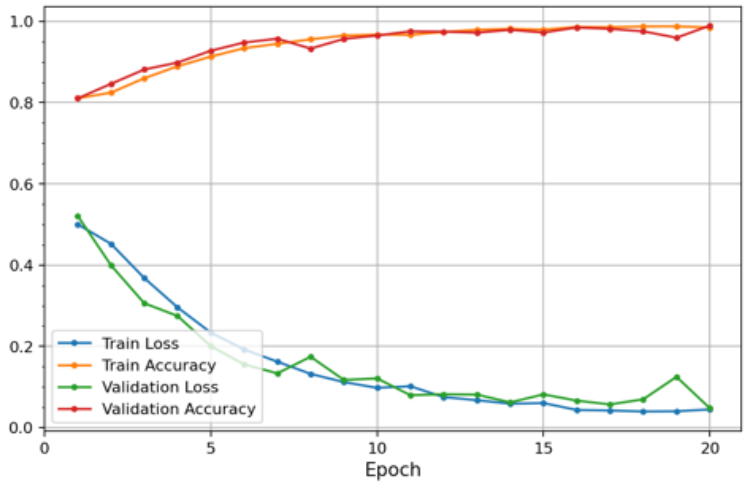


Fig. 5 - Training curves of ConViT-B (Attention Heads = 16) on 128x128 pixels resolution data

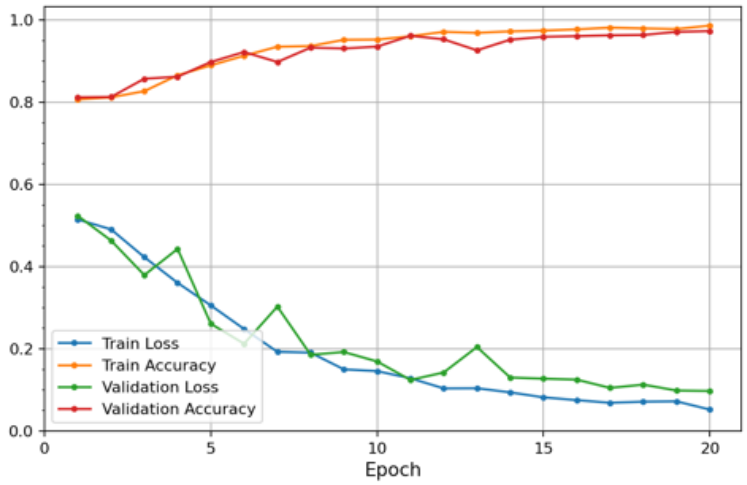


Fig. 6 - Training curves of ConViT-B (Attention Heads = 16) on 224x224 pixels resolution data

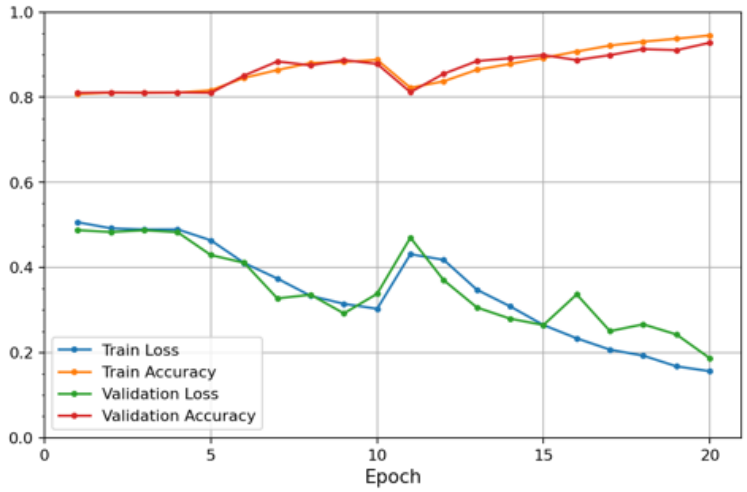


Fig. 7 - Training curves of ConViT-B (Attention Heads = 16) on 512x512 pixels resolution data

In Table 4, we compared the performance of the ConViT model with other state-of-the-art methods using the same dataset. Xception [21], Resnet50v2 [22] and Resnet50 [23] are convolutional neural networks which uses convolutional layers varied by their depth and width. Xception has 71 convolutional layers while Resnet50 and Resnet50v2 have 50 layers. It is to be noted that the study by Rahimzadeh et al. [10] uses the full dataset whereas study by Than et al. [20]

and this study uses only a subset of the dataset. This is due to computational limitation at time of experiments. However, the experiments results can still be used to compare with other methods. This is because usually, the performance of the model will increase when trained on a larger dataset.

Table 4 - Performance comparison with other state-of-the-art methods using same dataset

Methods	Accuracy (%)	Sensitivity (%)
Xception [10]	96.55	98.02
Resnet50v2 [10]	96.55	98.02
Resnet50 + Feature Pyramid [10]	98.49	94.96
ViT [20]	95.36	83.00
This Study (ConViT-B)	98.01	90.83

It is observed that ConViT-B with accuracy of 98.01% falls short by 0.48% and sensitivity of 90.83% falls short by 4.13% of the ResNet-50 with Feature Pyramid method. However, it is also to be noted that the ResNet-50 with Feature Pyramid method [10] used a transfer learning approach while the model used in this study (ConViT-B) is not pretrained. Besides, the study used the whole available dataset while in this study only partial of the dataset is used. A direct comparison cannot be made between the studies due to the difference of dataset sizes. However, the experiments in study can still offer meaningful comparisons. A reason for this is that although the dataset used is smaller, it is relatively large at 12059 images. We hypothesise with a larger training dataset, the performance of the model will increase.

Other than that, it is encouraging to see a huge performance rise from the previous study [20] which uses ViT that has accuracy of 95.36% and sensitivity of 83%. This shows that the combination of convolutions and transformer, can successfully learn the features of the two classes using fewer dataset as both global and local information of the images are covered and learned by the model. One trade-off of the introduction of GPSA layer in ViT is that ConViT will demand more computational resources, thus resulting in longer training time than ViT, in which we take as a reasonable trade-off.

5. Conclusion

In this study, we have investigated the effects of image pixels resolution and the number of attention heads used in ConViT on COVID-19 lung image classification. A peak accuracy of 98.01% was achieved by using 128x128 image pixels resolution and 16 attention heads in ConViT. This is further supported by other measures such as sensitivity, specificity, PPV, NPV and F1-score. Besides, the ConViT model has also achieved better performance than most of the other state-of-the-art methods that are using the same dataset, except for the use of Resnet50 + Feature Pyramid [10], with only 0.48% higher accuracy than ConViT. This proves that by introducing convolution to vision transformers, resulting in ConViT, the model performance increases. In short, this study has shown that the ConViT model can play a meaningful role to complement RT-PCR test on COVID-19 close contacts and patients with good accuracy and effectiveness. In future work, we plan to explore more on ConViT and its implementation on medical image datasets. This includes using a pretrained model, tuning hyperparameters, using a larger dataset etc. Besides, we would like to explore the processing speed of ConViT and compare it with other deep models. Lastly, we would also like to explore on providing explainability to the ConViT model used in this study for COVID-19 and other lung disease diagnosis tasks.

6. Acknowledgement

This work was supported by Ministry of Higher Education (MOHE) Malaysia, under the Fundamental Research Grant Scheme, (FRGS/1/2020/ICT02/SWIN/03/2, Project ID: 19009), project title “Explainable Artificial Intelligence (XAI) for Lung Disease Diagnosis and Severity Classification using Deep Learning”.

References

[1] ‘WHO Coronavirus (COVID-19) Dashboard’, *World Health Organization*. https://covid19.who.int/ (accessed Feb. 21, 2022).

[2] ‘Tracking SARS-CoV-2 variants’, *World Health Organization*. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed Feb. 21, 2022).

[3] A. Scohy, A. Anantharajah, M. Bodéus, B. Kabamba-Mukadi, A. Verroken, and H. Rodriguez-Villalobos, ‘Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis’, *Journal of Clinical Virology*, vol. 129, p. 104455, Aug. 2020, doi: 10.1016/j.jcv.2020.104455.

[4] A. Tahamtan and A. Ardebili, ‘Real-time RT-PCR in COVID-19 detection: issues affecting the results’, *Expert Review of Molecular Diagnostics*, vol. 20, no. 5, pp. 453-454, May 2020, doi: 10.1080/14737159.2020.1757437.

[5] K. Munne, V. Bhanothu, V. Bhor, V. Patel, S. D. Mahale, and S. Pande, 'Detection of SARS-CoV-2 infection by RT-PCR test: factors influencing interpretation of results', *VirusDisease*, vol. 32, no. 2, pp. 187-189, Jun. 2021, doi: 10.1007/s13337-021-00692-5.
[6] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, 'Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study', *The Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425-434, Apr. 2020, doi: 10.1016/S1473-3099(20)30086-4.
[7] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, T. M. L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, and others, 'Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT', *Radiology*, vol. 296, no. 2, pp. E46--E54, 2020.
[8] S. D'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, 'ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases', 2021.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, others, J. Uszkoreit, and N. Houlsby, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', *arXiv preprint arXiv:2010.11929*, Oct. 2020.
[10] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, 'A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset', *Biomedical Signal Processing and Control*, vol. 68, no. March, p. 102588, 2021, doi: 10.1016/j.bspc.2021.102588.
[11] E. Matsuyama, 'A Deep Learning Interpretable Model for Novel Coronavirus Disease (COVID-19) Screening with Chest CT Images', *Journal of Biomedical Science and Engineering*, vol. 13, no. 07, pp. 140-152, 2020, doi: 10.4236/jbise.2020.137014.
[12] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, 'Sample-efficient deep learning for COVID-19 diagnosis based on CT scans', *IEEE Transactions on Medical Imaging*, vol. XX, no. Xx, 2020, doi: 10.1101/2020.04.13.20063941.
[13] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, 'xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography', *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1-10, 2022, doi: 10.1109/JTEHM.2021.3134096.
[14] H. Chefer, S. Gur, and L. Wolf, 'Transformer Interpretability Beyond Attention Visualization', Dec. 2020.
[15] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, 'Vision Transformer using Low-level Chest X-ray Feature Corpus for COVID-19 Diagnosis and Severity Quantification', *arXiv preprint arXiv:2104.07235*, Apr. 2021.
[16] X. Gao, Y. Qian, and A. Gao, 'COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models', *arXiv preprint arXiv:2107.01682*, 2021.
[17] F. Mehboob, A. Rauf, R. Jiang, A. K. J. Saudagar, K. M. Malik, M. B. Khan, M. H. A. Hasnat, A. AlTameem, and M. AlKhathami, 'Towards robust diagnosis of COVID-19 using vision self-attention transformer', *Sci Rep*, vol. 12, no. 1, p. 8922, Dec. 2022, doi: 10.1038/s41598-022-13039-x.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, 'Attention is all you need', in *Advances in neural information processing systems*, 2017, pp. 5998-6008.
[19] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, 'Learning deep transformer models for machine translation', *arXiv preprint arXiv:1906.01787*, 2019.
[20] J. C. M. Than, P. L. Thon, O. M. Rijal, R. M. Kassim, A. Yunus, N. M. Noor, and P. Then, 'Preliminary Study on Patch Sizes in Vision Transformers (ViT) for COVID-19 and Diseased Lungs Classification', in *2021 IEEE National Biomedical Engineering Conference (NBEC)*, Nov. 2021, pp. 146-150. doi: 10.1109/NBEC53282.2021.9618751.
[21] F. Chollet, 'Xception: Deep Learning with Depthwise Separable Convolutions', Oct. 2016.
[22] K. He, X. Zhang, S. Ren, and J. Sun, 'Identity Mappings in Deep Residual Networks', Mar. 2016.
[23] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90.