Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Classification of COVID-19 and lung opacity using vision transformer on chest x-ray images

To cite this article: Manoochehr Noghanian Toroghi *et al* 2023 *J. Phys.: Conf. Ser.* **2622** 012016

Classification of COVID-19 and lung opacity using vision transformer on chest x-ray images

Manoochehr Noghanian Toroghi[1], Usman Ullah Sheikh[1*], Shima Shahi Irani[2]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor, Malaysia
[2]Department of Electrical Engineering, University of Applied Science, Bonn Rhein Sieg, Germany

Email: usman@fke.utm.my

Abstract. There are several recent works which had proposed an automatic computer-aided diagnosis (CAD) deep learning (DL) model to diagnose coronavirus disease 2019 (COVID-19) using chest x-ray images (CXR) to propose a high-accuracy CAD method to detect COVID-19 automatically. In this study, seven different models including Convolutional Neural Network (CNN) models such as VGG-16 and vision transformer (ViT) models, are proposed. The different proposed models are trained with a three-class balanced dataset consisting of 3,000 CXR images consisting of 1,000 CXR images for each class of COVID-19, Normal, and Lung-Opacity. A publicly available dataset to train and test the models is used from Kaggle-COVID-19-Radiography-Dataset. From the experiments, the accuracy of the VGG16 model is 93.44% and ViT's is 92.33%. Besides, the binary classification between two classes of COVID-19 and Normal CXR with a limited number of just 100 images for each class, using a transfer learning technique, with a validation accuracy of 97.5% is proposed.

Keywords— COVID-19, Convolutional Neural Network (CNN), VGG-16, Transfer Learning, Vision Transformer (ViT).

1. Introduction

A novel coronavirus, the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2, 2019-nCoV), was responsible for an acute typical respiratory disease in Wuhan, China in December 2019. Later it was recognized that human-to-human transmission played a major role in the subsequent outbreak [1]. To perform a reverse transcription polymerase chain reaction (RT-PCR) test, it is necessary to take a sample from the person requesting the test, and this operation must be done in person and by face-to-face contact between the applicant and the medical staff. Considering the high spreading potential of the coronavirus and general problems in the usual diagnosing test methods of COVID-19, diagnosing COVID-19 using medical images seems sensible. Diagnosis of specific diseases from a medical image, as input of a machine learning model, is categorized as an image classification problem in which the trained model will be able to classify any new input image because of the prediction of the model for diagnosing between different classes of the dataset by which the model has been trained.

CNN-based models are the most popular approach in solving CXR classification problems but achieved low accuracy in recent works [10] despite using some relatively sophisticated models. A more updated method which has recently been used in image classification is called the Vision Transformer (ViT). The transformer model is a simple yet scalable approach that can be applied to any kind of data if it is modulated as a sequence of embedding. In this paper, two different ViT models are proposed named the 'vit_base_resnet50_224_in21k' and the 'vit_base_patch16_224_in21k', which has a simpler structure. All these models are trained using the three-class dataset including COVID-19, Normal, and Lung-Opacity classes of CXR images. In addition, two other experiments were conducted in this study

Journal of Physics: Conference Series **2622** (2023) 012016 doi:10.1088/1742-6596/2622/1/012016

to perform binary classification of COVID-19 and normal chest images to test the ability of the transfer learning technique to solve image classification problems with limited datasets. In these two experiments, the transfer learning technique based on the pre-trained VGG-16 model was used to train the model using datasets of CXR images. Finally, the results of all the experiments proposed in this study are compared.

2. Related Works

Recent research works have tried to propose different models to diagnose COVID-19 using CXR and DL models, because of the ability to extract different features from training images using different convolutional filters. CNNs were first introduced by [5] in 1998 and used for digit recognition. In [3], the authors introduced a transfer learning CNN model, based on the pre-trained VGG16, to classify three classes of COVID-19, Normal, and Other Pneumonia from CXR images with an overall accuracy of 94.5, 98.4 % sensitivity and 98.0 % specificity in classifying cases with and without COVID-19 infection. The F1-score in their study for classes of Other Pneumonia, Normal, and COVID-19 is 0.96, 0.93, and 0.84 respectively. When compared to other models such as ResNet50, MobileNetV2, M-Inception, COVID-Net, CoroNet (Xception), their VGG16 proposed model outperformed others in both binary and three-class classification. In addition, their model achieved noticeably higher performance in the case of using CXR compared to CT images. In [8], the authors have proposed a CNN architecture model to classify COVID-19 using CXR images and showed the importance of selecting the correct model with the appropriate number of CNN layers. They have provided a comparison between the results of a classification in three different CNN models in which their CNN had 3, 4, and 5 layers of convolution with max-pooling namely models 1, 2, and 3 respectively. They used a small dataset of 330 CXR images which are equally divided into two classes: COVID-19 and Normal, for training the model. Similarly, an equally divided image set of 82 CXR was used for validating the model. Figure 1 depicts their proposed sequential CNN, which is one of the models proposed in this paper.

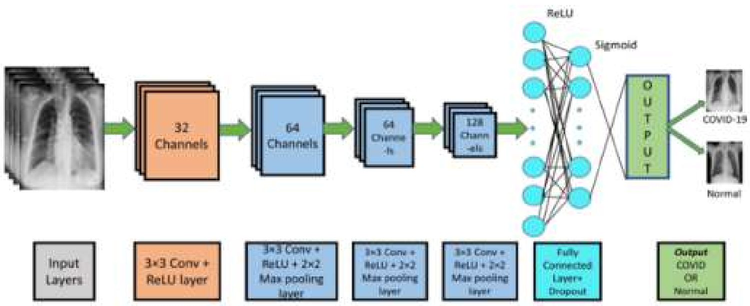

Figure 1. CNN architecture proposed in [8]

Model 1, in their study, performed with accuracy and precision of 97.56% and 95.34% respectively. Moreover, this model is compared to two CNN models with a different number of convolutional layers. The comparative studies show a better F1-score and overall performance of Model 1 than Model 2. This model can be further improved with the availability of a larger dataset. Hence, CNN has great prospects in detecting COVID-19 with very limited time, resources, and costs. In [7], the authors proposed a model to detect and classify tuberculosis (TB) disease in addition to the other five different diseases. Their model used the transfer learning technique based on a pre-trained VGG Net with three different learning rates. Their model reached an average AUC of 0.908 across the six lung diseases.

From these research works, it was shown that VGG-16 was the highest-performing DL model in image classification, but it does not mean using relatively complicated DL models gives high accuracy in image classification in general. For instance, in [10] the authors proposed a comparison table of classification performance obtained from different pre-trained CNN models to classify from a three-class dataset including Normal, COVID-19, and Pneumonia. In their study, there is low accuracy of some models such as MobileNetV2 with total accuracy of 39.7 % and DensNet201 with 38.23 % in their specific image classification problem.

However, in their study, some models like VGG-16 gave a total accuracy of 95.88 %. In this paper, a CNN-based VGG-16 architecture model as well as a transfer learning technique, has been considered as the starting point in three-class classification. Then, the ViT image classification model has been used as a more updated classifier method.

3. CNN-based Models

CNN-based models are the most popular deep learning models in solving image classification problems. The reason is more related to the ability of CNNs to extract the different features of an input image using different convolutional filters. The models can be trained as below:

- **Training from scratch**
  VGG16 model is a CNN network trained on a subset of the "ImageNet" dataset, a collection of over 14 million images belonging to 22,000 categories. In the ImageNet Classification Challenge in 2014, VGG16 achieved 92.7% classification accuracy. A simple CNN model and a model based on the general architecture of VGG16 with appropriate hyperparameters were proposed in this study, both trained from scratch.

- **Transfer learning**
  Transfer learning is one of the most important techniques of deep learning. Instead of starting training data from scratch, it is possible to use a pre-trained model trained as a starting point to train the target model on a smaller dataset for a given task. It is a technique by which the network can be trained a lot faster with better results. As the name implies, transfer learning means transferring knowledge (feature maps) that a neural network has learned from being trained on a specific dataset to another related problem. One or more layers from the trained model are then used in a new model trained on the problem of interest. The pre-trained architecture of VGG16 can detect generic visual features present in the dataset and it is the next model proposed in this paper as both a feature extractor (with no fine-tuning) and as a fine tuner, separately.

4. ViT Models

Self-attention-based architectures, in particular Transformers [11], have become the model of choice in natural language processing and computer vision problems. A transformer in machine learning is a deep learning model that uses the mechanisms of attention, differentially weighing the significance of each part of the input data. Transformers in machine learning are composed of multiple self-attention layers. The Transformer Encoder consists of:

i. *Multi-Head Self Attention Layer* to concatenate the multiple attention outputs linearly to expected dimensions. The multiple attention heads help learn local and global dependencies in the image.

ii. *Multi-Layer Perceptron* contains two-layer with Gaussian Error Linear Unit.

iii. *Layer Norm* is applied before every block as it does not introduce any new dependencies between the training images. It helps improve the training time and generalization performance.

iv. *Residual connections* are applied after every block as they allow the gradients to flow through the network directly without passing through non-linear activations.

In 2021, Alexy et al. [9], introduced the Vision Transformer (ViT). ViT models are pre-trained transformer models for image processing tasks. The models are trained on ImageNet and ImageNet-21k datasets. ViT models outperform CNN models on recognition benchmarks such as ImageNet, CIFAR-100, and VTAB. As an overview of vision transformer in image classification, after splitting an image into patches, which have fixed sizes, each patch embeds linearly and is followed by adding position embedding. Then, the results, which are a sequence of vectors, are fed to a standard transformer encoder. It manipulates the input sequence with a multi-self-attention and embeds as much information as possible for classification into the classer token, as Figure 2 depicts. The self-attention layer in ViT makes it possible to embed information globally across the overall image. The model also learns from training data to encode the relative location of the image patches to reconstruct the structure of the image.

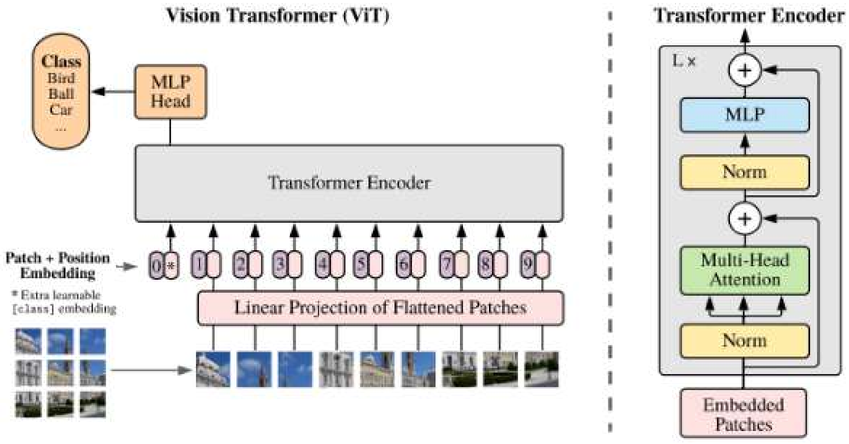

Figure 2. Vision Transformers (ViT) [9]

5. Dataset

The dataset used in this work is the Kaggle Radiography Dataset which consists of three classes named COVID-19, Normal, and Lung Opacity. The dataset contains 3,616 COVID-19-positive cases, 10,192 normal CXR and 6,012 with Lung Opacity (non-COVID lung infection). For training and testing, a balanced dataset is prepared for this study consisting of 1,000 images for each class separated into 80% (800 images) for training and 20% (200 images) for validation. Figure 3 shows samples of CXR from the dataset for each class. In this study, seven models are experimented with.

Model 1: Three-class (COVID-Normal-Lung Opacity) classifier using CNN.
Model 2: Three-class (COVID-Normal-Lung Opacity) classifier using VGG16.
Model 3: Three-class (COVID-Normal-Lung Opacity) classifier using pre-trained VGG-16 without fine tuning.
Model 4: Three-class (COVID-Normal-Lung Opacity) classifier using pre-trained VGG-16 with fine tuning.
Model 5: Two-class (COVID-Normal) (Each class includes 200 images = small dataset). Classifier by transfer learning using pre-trained VGG-16.
Model 6: Two-class (COVI-Normal) (Each class includes 1,000 images). Classifier by transfer learning using pre-trained VGG-16.
Model 7: Three-class (COVID-Normal-Lung Opacity) classifier using ‘vit_base_resnet50_224_in21k’ ViT model.
Model 8: Three-class (COVID-Normal-Lung Opacity) classifier using ‘vit_base_patch16_224_in21k’ Vision Transformer Model.

6. Results

Table 1 shows the results of seven different models proposed in this study as well as the models' parameters.

Table 1. Comparison table of the results

Model	Classes	Dataset Size	Model Parameters	Accuracy (%)	Runtime
Model 1 CNN	Normal, COVID-19, Lung Opacity	3,000	Image size: 56×56 Optimizer: Adam, Layers: 7	91.1 %	-
Model 2 VGG16	Normal, COVID-19, Lung Opacity	3,000	Image size: 224×224 Optimizer: Adam, Layers: 16, LR=0.001	93.44 %	-
Model 3 VGG16 - Without fine tuning	Normal, COVID-19, Lung Opacity	3,000	Image size: 224×224 Optimizer: Adam Layers: 16 Fine tune: 0, LR:0.0001	88.8 %	17 m:36s
Model 4 VGG16 - With fine tuning	Normal, COVID-19, Lung Opacity	3,000	Image size: 224×224 Optimizer: Adam, Layers: 16, Fine tune=2, LR:0.0001	90.5 %	18 m:52s
Model 5 Pre-trained VGG16 (Transfer Learning)	Normal, COVID-19	400	Image size: 224×224 Optimizer: Adam, LR: 0.001 Pool size: (4,4), Dropout: 0.5	100.0 %	-
Model 6 Pre-trained VGG16 (Transfer Learning)	Normal, COVID-19	2,000	Image size: 224×224 Optimizer: Adam, LR: 0.001 Pool size: (4,4), Dropout: 0.5	97.5 %	-
Model 7 'vit_base_resnet50_224_in21k'	Normal, COVID-19, Lung Opacity	3,000	Image size: 128×128 Optimizer: SGD Transformer resize:150 LR:0.01	83.0 %	18 m:15s
Model 8 'vit_base_patch16_224_in21k'	Normal, COVID-19, Lung Opacity	3,000	Image size: 224×224 Optimizer: Adam Transformer resize:150 LR: 0.001	92.3 %	14 m:24s

Journal of Physics: Conference Series **2622** (2023) 012016 doi:10.1088/1742-6596/2622/1/012016

7. Conclusion

In conclusion, CNN is a powerful tool to extract the specific features of an image, as shown in the model's performance (91.1%) accuracy. VGG-16 as a CNN-based deep learning model has a good performance in solving image classification problems. Using a pre-trained VGG-16 model with fine-tuning can improve the accuracy, (88.8% and 90.5%). Using position embedding in the encoder-decoder architecture of ViT can keep the location of the image's features. The lower accuracy of ViT models compared to CNN-based models in this study is expected because of the small and medium size of the datasets. ViT outperforms CNN when training is based on large datasets.

(a)

(b)

(c)

Figure 3. Sample images from the Kaggle Radiography Dataset, (a) COVID-19, (b) Lung Opacity and (c) Normal CXR.

Acknowledgement

The authors would like to thank the Ministry of Higher Education Malaysia (KPT) and Universiti Teknologi Malaysia (UTM) for their support under the UTM Fundamental Research Grant (UTMFR), grant number Q.J130000.3823.22H29.

References

[1] Chowdhury, Nihad-Rahman, Md Muhtadir-Kabir, Ashad, S.-H. (2020): '*A Parallel-Dilated Convolutional Neural Network Architecture for Detecting COVID-19 from Chest X- Ray Images*,', DOI,10.31224/osf.io/myp6c

[2] Jain, Govardhan- Mittal, Deepti -Thakur, Daksh- Mittal, MadhupK.,(2020) '*A deep learning approach to detect Covid-19 coronavirus with X-Ray images*', Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier B.V., 2020, DOI:10.1016/j.bbe.2020.08.008

[3] Heidari, Morteza Mirniaharikandehei, Seyedehnafiseh Khuzani, Abolfazl ZargariDanala, Gopichandh Qiu, Yuchen Zheng, Bin – (2020) '*Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms.', International* Journal of Medical Informatics, 2020' DOI: 10.1016/j.ijmedinf.2020.104284

[4] Islam, Md Zabirul-Islam, Md Milon-Asraf, Amanullah (2020) '*A combined Deep CNN-LSTM network for the detection of novel Coronavirus (COVID19) using X-ray images*', Informatics in Medicine Unlocked, DOI:10.1016/j.imu.2020.100412

[5] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998) '*Gradient-based learning applied to document recognition*,', IEEE.

[6] Ze Liu and Yutong Lin and Yue Cao and Han Hu, Yixuan We and Zheng Zhang Stephen Lin Baining Guo, (2021) '*Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*.', arXiv:2103.14030v1 [cs.CV] 25 Mar 2021.

[7] Hasan Nabeel Saleem, Usman Ullah Sheikh, and Saifulnizam Abd. Khalid, (2021) '*Classification of Chest Diseases from X-ray Images on the CheXpert Dataset*.' DOI: 10.1007/978-981-16-0749-3_64

[8] Khandaker Foysal Haque, Fatin Farhan Haque, Lisa Gandy, and Ahmed Abdelgawad, (2020) '*Automatic Detection of COVID-19 from Chest X-ray Images with Convolutional Neural Networks*', International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2020, pp. 125-130, doi: 10.1109/iCCECE49321.2020.9231235.

[9] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, (2021) '*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*', arXiv 2010.11929 125-130, doi: 10.1109/iCCECE49321.2020.9231235.

[10] Konstantinos Tserpes, Antonios Makris and Ioannis Kontopoulos, (2020) '*COVID-19 detection from chest X-Ray images using Deep Learning and Convolutional Neural Networks*', DOI: 10.1145/3411408.3411416.

[11] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, (2017) '*Attention Is All You Need*', arXiv doi: 10.48550/ARXIV.1706.03762