A transformer-based framework for automatic COVID19 diagnosis in chest CTs

Lei Zhang
Laboratory of Vision Engineering (LoVE)
University of Lincoln
lzhang@lincoln.ac.uk

Yan Wen
Laboratory of Vision Engineering (LoVE)
University of Lincoln
ywen@lincoln.ac.uk

Abstract

Automated diagnosis of covid19 in chest CTs is becoming a clinically important technique to support precision and efficient diagnosis and treatment planning. A few efforts have been made to automatically diagnose the COVID-19 in CTs using CNNs, and the task still remains a challenge. In this paper, we present a transformer-based framework for COVID19 classification. We attempt to expand the adaption of vision transformer as a robust feature learner to the 3D CTs to diagnose the COVID-19. The framework consists of two main stages: lung segmentation using UNet followed by the classification, in which the features extracted from each CT slice using Swin transformer in a CT scan are aggregated into 3D volume level feature. We also investigated the performance of using the robust CNNs (BiT and EfficientNetV2) as backbones in the framework. The dataset from the ICCV workshop: MIA-COV19D, is used in our experiments. The evaluation results show that the method with the backbone of Swin transformer gain the best F1 score of 0.935 on the validation dataset, while the CNN based backbone of EfficientNetV2 has the competitive classification performance with the best precision of 93.7%. The final prediction model with Swin transformer achieves the F1 score of 0.84 on the test dataset, which doesn't require an additional post-processing stage.

1. Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic has led to tremendous public health concern across the world. The COVID-19 caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) harms patients due to the lack of effective clinical treatment with antiviral drugs [4], though some vaccines are available to date. As such, it is of great crucial to prevent the further spread of the disease by introducing a faster and more efficient diagnosis of COVID-19, where it's been commonly accepted that the Reverse transcription-polymerase Chain Reaction (RT-PCR) test is the gold standard for the diagnosis of the COVID-19 [1]. However, the high false-negative rate of RT-PCR raises a concern of missing some potential cases [4]. Therefore, as an important complementary tool to the RT-PCR [1, 7], chest CT imaging has been used in clinical flow to diagnose and grade the COVID-19. Studies [20, 8] have reported that some radiographic signs can be observed in chest CT images, such as ground-glass opacities (GGO), crazy-paving pattern and consolidation. However, with the rapidly increasing number of patients in the current situation, it leads to a great challenge for manually interpreting CTs by radiologists. For instance, lesion/GGOs presented in CTs are tiny, especially in early-stage cases that could substantially increase the missed detection rate. Moreover, the diagnosis error could be introduced due to the similar signs between COVID-19 and Non-COVID-19 CTs [22]. In this case, automated interpretation of chest CT is non-trivial and becoming a clinically important technique to support precision and efficient diagnosis, treatment planning and quantify the discharge criteria [20]. Recently, a number of deep learning-based methods have been proposed for the automatic diagnosis of COVID-19. For instance, Wang [26] proposed a 3D convolutional neural network (CNN) for COVID-19 classification and localization, where the diagnosis of the COVID19 is treated as a typical image classification problem using a 3D CNN, the lesion localization was achieved by employing the class activation mapping (CAM). A specific CNN was designed names COVIDNet-CT presented in the studies [9, 2], which was further enhanced to the COVIDNet-CT2 with fewer parameters. The core architecture in these methods was constructed via machine-driven design exploration. Li *et al*. [16] proposed a 2D CNN for feature extraction from each slice in a CT, and the slice-level features are further fused via a max-pooling layer. In the study [22], the attention mechanism (including both channel-wise attention (CA) and depth-wise attention (DA)) was integrated into a modified 3D Resnet18 that leverages a residual network to automatically identify COVID-19 from other common pneumonia and normal people in the chest CTs. A unified DNNs based framework presented

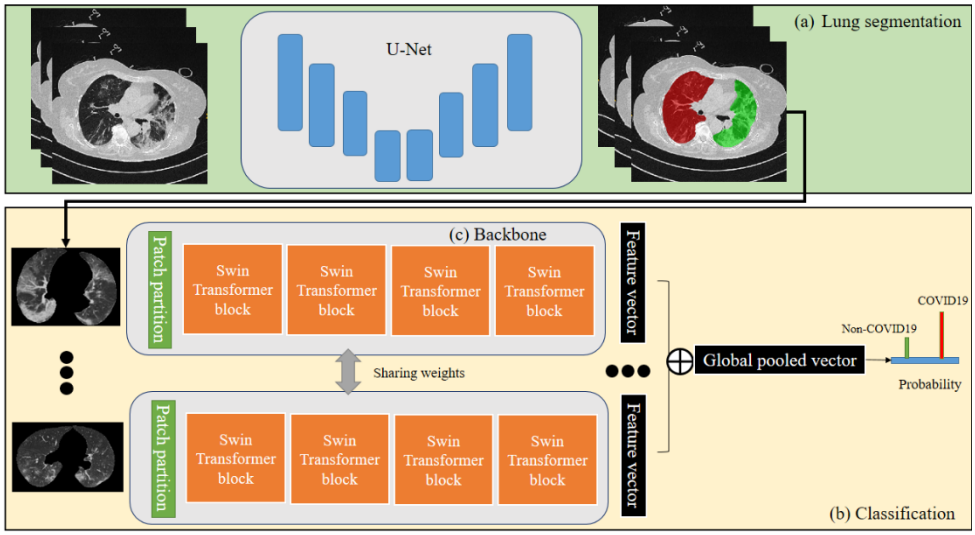

Figure 1. The general framework of our covid19 diagnosis method. (a) is the pre-processing stage for lung segmentation, (b) is the image classification stage, where the features extracted from each slide in a CT scan are aggregated into CT volume level feature via max-pooling.

in the studies [12, 14, 13, 15], which couples the CNNs and RNNs for prediction of COVID-19 and pneumonia from CT scans. Domain adaption of applying the unsupervised latent representation learning allows the method to be portable in other independent data without extensive training. Although plenty of efforts have been made to automatically diagnose the COVID-19 in CTs using deep learning, the task remains a challenge due to similarities of appearances between COVID and Non-COVID CTs. Moreover, almost all existing methods heavily rely on the strong capacity of feature representation learned by CNN. However, few attempts have been made to adopt the Vison transformer architectures [6, 18], though the latest studies [3, 21] have shown its robustness and superior performance in image classification compared to the CNN based architectures. In this paper, we seek to expand the applicability of vision transformer such that it can be served as a more robust learner for diagnosis of the COVID-19 in chest CTs. We conducted our study by participating in the ICCV21 workshop of "AI-enabled Medical Image Analysis Workshop and Covid-19 Diagnosis Competition (MIA-COV19D)" and reported our method and experimental results in this paper.

2. Methods and Materials

2.1. Methods

In this study, we introduce a vision transformer classification network for COVID19 diagnosis in chest CTs, inspired by the success of the Swin vision transformer and CT classification work in [26, 16]. The general framework is shown in Figure 1, which consists of 2 main stages: lung segmentation as pre-processing followed by the image classification using Swin transformer as backbone [18]. In the first stage, a pre-trained Unet [10] for lung segmentation in CTs was employed to generate lung masks. There are two merits to do so. 1) it allows the learning to be limited in the specific lung region. 2) while it could reduce the computational cost using the deep architectures for the CT volume. In the second stage, the Swin vision transformer is used to extract features from each 2D slices in a CT volume, which are further aggregated into volume level features via a max-pooling layer.

2.1.1 Backbone of swim transformer

Recently, the success of deploying the transformer with attention mechanism [6, 24] in the Computer Vision (CV) motivates the researchers to seek its applicability and adaption to various data modalities in different applications. We attempt to expand its adaption as a robust feature learner to the 3D CTs for the diagnosis of the COVID-19. More specifically, we adopted the swim transformer [18] as the backbone for the feature learning in the framework shown in Fig. 1, given the superior scaling strategy deployed compared to the previous ViT [6, 24]. In this section, we briefly introduce the preliminaries related to the ViT and its variant-the swim transformer. The ViT is constructed by stacking transformer blocks, consisting of two main components: Multi-head Self Attention (MSA) and Feed-Forward Network (FFN). In the image classification task, the images are first divided into patch embedding sequences and then fed into the transformer blocks.

Multi-head Self Attention (MSA) is the core component of the ViT model, which is essentially dot product at-

tention [25]. The self-attention is expressed as follows:

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

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$ represent Query, Key and Value matrices, respectively. Here, the W is a linear transformation with weighs w, and input X is computed from a sequence of N input vectors via linear projections. d is the query/key dimension, and $\sqrt{d}$ in (1) is a normalization term. The MSA, as the name implied, h self-attention (heads) are considered in the transformation that allows the more representative features to be learned through the stacked blocks. The output of each head forms a sequence of size $N \times d$. So that a sequence of $N \times dh$ derived from h heads is then fed through a linear transformation layer that produces a final output of the size of $N \times D$ from MSA.

Each **transformer block** comprises MSA, Layer Normalization (LN), FFN, and skip connection. And its implementation can be formulated as follows:

$$\begin{aligned} z_l^{'} &= MSA(LN(z_{l-1}) + z_{l-1}) \\ z_l &= FFN(LN(z_l^{'}) + z_l^{'}) \end{aligned} \quad (2)$$

where the $z_l^{'}$ and z_l are the output features of the MSA and FFN at the current lth block, respectively. The standard architectures [6, 24] equipped with the aforementioned transformer blocks are characterized by conducing the global self-attention that leads to quadratic complexity with respect to input sequence size. Alternatively, as a general-purpose architecture, **the Swin transformer** achieves a better speed-accuracy trade-off compared to the standard ViTs. It is a hierarchical transformer that is built by replacing the standard MSA in the transformer block by the shifted window. The code design of shifted window enables to compute self-attention within local windows, instead of learning representation with self-attention globally in the standard ViTs. To overcome the lack of connection across the non-overlapped windows generated using a regular window partitioning strategy in the design, a shifted window partitioning method is proposed, which adopts a windowing strategy that shifted from the regularly partitioned windows in the preceding layer. As such, following the expresses (2), the consecutive Swin transformer blocks are computed as

$$\begin{aligned} z_l^{'} &= W\_MSA(LN(z_{l-1}) + z_{l-1}) \\ z_l &= FFN(LN(z_l^{'}) + z_l^{'}) \\ z_{l+1}^{'} &= SW\_MSA(LN(z_l) + z_l) \\ z_{l+1}^{'} &= FFN(LN(z_{l+1}^{'}) + z_{l+1}^{'}) \end{aligned} \quad (3)$$

Where W_MSA is window based MSA using regularly windowing configuration, and the SW_MSA is windowing using shifted window partitioning configuration, To provide a comparison between Transformer architecture and CNN based architecture, we replaced the backbone shown in the (c) of Figure 1 by the BiT and EfficientNetV2.(we refer the reader to [11, 23] for more details on these two state-of-the-art CNN models).

2.1.2 Evaluation methods

We utilize traditional F1 score for binary classification to measure the performance among the different settings for model selection.

$$F1_{score} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \quad (4)$$

Finally, our method is evaluated by Macro F1 Score for competitive comparison, which is the challenge preferred metrics.

$$Macro\ F1_{score} = \frac{1}{n}\sum_{i=0}^{N} F1_{score_i} \quad (5)$$

Where i is the class/label index and N is number of classes/labels.

2.2. Dataset

We use the MIA-COV19D competition dataset [12] in our experiments, which splits into training, validation and test sets. The dataset consists of about 5,000 CT scan series, and each scan consists of a sequence of 50-700 2-D CT slices, where the ground truths are provided with respect to labels of Covid-19/non-Covid-19 diagnosis for each case. The training set contains 1560 CT scans which include 690 COVID-19 cases and 870 Non-COVID-19 cases. The validation set consists of 374 CT scans, of which 165 are COVID-19 cases, and 209 are non-COVID-19 cases. No additional datasets were used in our experiments for CTs classification. The test dataset was used to validate our method which contains additional 3455 CT scans.

3. Experiments and results

3.1. Data preprocessing

Each CT scan consists of a sequence of 2D CT slices, and each slice is store in jpeg format. Due to the fact that pre-trained weights of the Unet [10] for lung segmentations were obtained by training and evaluating from the data in the standard Hounsfield scale. [17] We projected the image intensity back to the Hounsfield scale that allows us to produce consistent lung segmentation masks using the pre-trained weights. To further improve the image loading efficiency in the training, a sequence of slides in a case is converted to a 3D CT volume saved in NIfTI format.

Backbone	Input size	Key architecture	Parameters
BigTransfer	224×224	BiT-M-ResNetV2_101×1	45m
EfficientNetv2	224×224	EfficientNetV2-M	55m
SwinTransformer	224×224	Swin-B	88m

Table 1. The experimental settings with different backbones

Backbone	Precision(%)	Recall(%)	Accuracy(%)	F1(COVID) score
BiT-M	90.1	91.3	91.8	0.927
EfficientNetV2-M	93.7	92.5	94.0	0.931
Swin-B	93.2	93.8	94.3	**0.935**

Table 2. Evaluation results on validation dataset and comparisons

In our experiments, only lung regions generated via lung masks(segmentations) are taken into account in classification to reduce the computational cost. In addition, each slide in a CT is rescaled considering the limitation of GPU memory and comparability for different backbones.

3.2. Experiments and Settings

In order to explore the applicability of the Swin transformer for the COVID19 classification in chest CTs, we conducted several experiments with the settings of using different backbones in the framework (Figure 1). More specifically, the deployed backbones include Swin transformer (Swin-B), BigTranfer (BiT-M) and EfficientNetV2 (EfficientNetV2-M). The CNN-based representations learners are provided mainly for comparisons. Table 1 summarises the used settings of these backbones. We considered the speed-accuracy trade-off and size of networks due to the limitation of GPU memory when we selected architectures within different backbones. The image size for all architectures remains the same, with the size of 224×224 pixels.

Our experiments are implemented on the PC with specifications of Intel Core i9 10980XE Processor, The GeForce RTX 3090 with Memory Size of 24 GB.

3.3. Training protocols

For all architectures in the training stage, we employed the same strategy. Namely, we fine-tunned the weights based on the pre-trained weights of the ImageNet-21k dataset [5]. For the Swin transformer backbone, we following the same protocol used in the fine-tuning stage in [18]. We employed the AdamW [19] optimizer with an initial learning rate of 10-5, weight decay of 0.05. For both CNN based networks, we employed the Adam with the learning rate of 10-5. For all training, we set in a total of 50 training epochs, early stopping training strategy is adopted, of which condition was set as the no decrease of the validation loss within ten epochs. Our implementations are open-sourced and available at https://github.com/leizhangtech/COVID19T.

Dataset	Methods	Macro F1
Validation	Baseline-ResNet-GRU [12]	0.70
Validation	Our method	**0.94**
Test	Baseline-ResNet-GRU	0.67
Test	Our method	0.84

Table 3. Evaluation results on validation dataset and comparisons

3.4. Experimental results

The experimental results of the validation dataset are summarized in Table 2. The evaluation metrics include the precision, recall, accuracy and F1 score. Three backbones described in Table1 were trained on the training dataset, and the trained models are applied to the validation dataset. We can observe from Table 2 that the method with the backbone of Swin-B has the best F1 of 0.935, which shows its robust feature learning capability. Meanwhile, we can see that the backbone of EfficientNetV2-M has competitive performance in term of F1 and accuracy and gain the best precision of 93.7%.

We also compare our method to the baseline method in terms of F1 score on both validation and test datasets shown in Table 3. The best model obtained from architecture with the Swin-B backbone was selected as the final prediction model on the testing dataset, as it shows the best F1 score over the validation dataset. However, it's worth noting that the framework equipped with the backbone of EfficientNetV2-M achieves a good speed-accuracy trade-off according to the results on the validation dataset. This suggests a potential of improvement for the classification could be achieved by simply increasing the model size in future work.

4. Conclusion

This paper presents a deep learning-based general framework for the diagnosis of COVID19 using chest CTs. The

framework consists of two stages: Unet based lung segmentation followed by the image classification with Swin transformer backbone. Our results show that the framework with the backbone of the Swin-B gains the best classification performance with a 0.935 of F1 score. This reflects the strong applicability of vision transformer that can be served as a robust learner for diagnosis of the COVID-19 in chest CTs. It was not surprising that the CNN base backbone using EfficientV2 has a competitive performance with fewer parameters than the Swin-B.

References

[1] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, 296(2):E32–E40, 2020. 1

[2] Anirudh Ambati and Shiv Ram Dubey. Ac-covidnet: Attention guided contrastive cnn for recognition of covid-19 in chest x-ray images. *arXiv preprint arXiv:2105.10239*, 2021. 1

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *arXiv preprint arXiv:2103.14586*, 2021. 2

[4] Jasper Fuk-Woo Chan, Shuofeng Yuan, Kin-Hang Kok, Kelvin Kai-Wang To, Hin Chu, Jin Yang, Fanfan Xing, Jieling Liu, Cyril Chik-Yan Yip, Rosana Wing-Shan Poon, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The lancet*, 395(10223):514–523, 2020. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[7] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology*, 296(2):E115–E117, 2020. 1

[8] Leiwen Fu, Bingyi Wang, Tanwei Yuan, Xiaoting Chen, Yunlong Ao, Thomas Fitzpatrick, Peiyang Li, Yiguo Zhou, Yi-fan Lin, Qibin Duan, et al. Clinical characteristics of coronavirus disease 2019 (covid-19) in china: a systematic review and meta-analysis. *Journal of Infection*, 80(6):656–665, 2020. 1

[9] Hayden Gunraj, 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. 1

[10] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020. 2, 3

[11] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 3

[12] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021. 2, 3, 4

[13] Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020. 2

[14] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018. 2

[15] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, pages 251–267, 2020. 2

[16] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: evaluation of the diagnostic accuracy. *Radiology*, 296(2):E65–E71, 2020. 1, 2

[17] Aleksandra Lis, Pawel Sniatala, and Marek Konkol. Computer-aided diagnosis in lungs radiography. In *2018 25th International Conference" Mixed Design of Integrated Circuits and System"(MIXDES)*, pages 427–430. IEEE, 2018. 3

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 4

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[20] Feng Pan, Tianhe Ye, Peng Sun, Shan Gui, Bo Liang, Lingli Li, Dandan Zheng, Jiazheng Wang, Richard L Hesketh, Lian Yang, et al. Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia. *Radiology*, 2020. 1

[21] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 2

[22] Jun Shi, Huite Yi, Xiaoyu Hao, Hong An, and Wei Wei. Dual-attention residual network for automatic diagnosis of covid-19. *arXiv preprint arXiv:2105.06779*, 2021. 1

[23] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021. 3

[24] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2, 3

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[26] Xinggang Wang, Xianbo Deng, Qing Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, and Chuansheng Zheng. A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. *IEEE transactions on medical imaging*, 39(8):2615–2625, 2020. 1, 2