A fusing Transformer and CNN on Interpretable COVID-19 Detection

Zhuohui Pan[1a], Yujuan Chen[2b*]

[1]School of Zhejiang University of Finance and Economics, China,
[2]School of Zhejiang University of Finance and Economics, China,

[a]987090710@zufe.edu.cn
[b]chenyj@zufe.edu.cn

Abstract. Although computer-aided diagnosis has become an important tool for rapid detection of lung diseases, the reliability of algorithm visualization on chest X-ray (CXR) images remains a challenge. This study explores the detection performance of a fusion model combining Transformer and CNN models. A decision constraint module was designed to achieve interpretable pneumonia detection. The performance of the decision constraint module was observed using Grad-CAM technique, and experimental results demonstrate that it outperforms lung mask segmentation. By activating transfer learning, our parallel combination model effectively identifies COVID-19 categories with a test set accuracy of 98.65%.

Keywords: COVID-19, CNN, Transformer, Lung segmentation, Transfer Learning

1 Introduction

COVID-19 caused a global health crisis and is one of the deadliest pandemics of this century. While the virus has become less virulent, it remains highly transmissible. It is expected that the virus will continue to coexist with us in the foreseeable future. Therefore, early screening is essential to prevent further spread.

Medical image analysis is an essential tool and technical approach in medical research, clinical disease diagnosis, and treatment. It has gained significant momentum due to the continuous advancement of medical imaging technology and computer technology. The availability of open-source medical image databases further supports the development of this field. (Rajpurkar P al., 2017[1]; Wang X al., 2017 [2]) studied the early deep learning classification application of low-level features of lung diseases such as atelectasis, infiltration, pneumothorax, etc. on the chestx-ray14 dataset.

Convolutional neural networks (CNNs) have achieved unprecedented performance in many tasks such as medical image classification, localization and segmentation. The CNN-based approach has also been used for the recognition of X-ray images of COVID-19 by (Wang Wei al., 2021[3]; Yang Jiezhi al., 2021[4]; Yi Sanli al., 2021[5]). CNNs have achieved success but rely on vision's inductive bias and lack efficiency in

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_46

capturing global context. Stacking convolutional layers expands receptive fields for accessing distant information, but downsampling weakens low-level features and discards some as resolution decreases. Thus, researchers focus on a new framework using self-attention to model global context. (Jiang J al., 2021[6]; Krishnan K S., 2021[7]; shome D al., 2021[8]; Gao x al., 2021[9]) used Vision Transformer and Swin Transformer, which were variants of self-attention mechanisms. Currently, the extension of self-attention has been successfully applied to downstream tasks in computer vision.

Deep learning models have had success in decision-making but face challenges. Unlike support vector machines with mathematical derivations, deep learning optimizes through reward mechanisms and lacks interpretability on small datasets. This makes it difficult for evaluation metrics to reflect the model's understanding of pathology when medical data is limited. (Kerman d s al., 2018[10]) they validated the performance of CNN models on image tasks related to the eyeball and lung organs. Specifically, they trained on 5232 chest X-ray images, and obtained 92.8% accuracy when distinguishing pneumonia from normal chest X-ray on test set. While (Rahman t al., 2021[11]) explored the performance of different CNN models on 11825 chest X-ray images under multiple image enhancement technologies. They proposed using lung segmentation mask images as inputs to introduce a bias towards the lung region and enhance model interpretability. However, this led to a decrease in the accuracy of the best pneumonia identification model from 96.29% to 93.22%. In my opinion, despite the decrease in metrics, the visual results of the model trained on small data are more interpretable.

Finally, this paper proposes a combined model of parallel CNN and Transformer, with the decision constraint function as a separate plug-in module. The aim is to validate the model's learning ability through both metrics and visualization, as Fig 1.

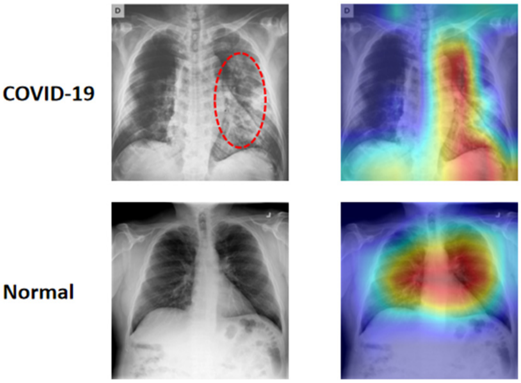

Fig. 1. Grad-CAM visualizes COVID-19 CXR and Normal CXR to show the learning position of CNN model

In summary, our main contributions are as follows.

- The concurrent structure is used to replace the serial feature pattern, which allows for better integration of local features and global semantics.

- The decision constraint module is established, which exhibits strong applicability and outperforms lung segmentation mask images as inputs.
- This paper goes beyond the pursuit of model classification accuracy and expands the traceability of decision-making basis using CAM technology to enhance the reliability of providing auxiliary diagnosis for clinicians.

2 Related works

2.1 Self-attention

The self-attention mechanism was proposed by the Transformer model. (Vaswani et al., 2017[12]), and was originally used to deal with machine translation tasks. The method used by VIT (Dosovitskiy A et al., 2020[13]) to introduce self-attention mechanism is by flattening the 2D image into a one-dimensional sequence and learning the representation power of the data through iterative encoding and decoding. While it breaks the spatial constraints of convolutional windows, it loses the translational invariance induction bias, resulting in an increase in model ceiling and the difficulty of fitting.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The vector representation of Q, K, and V here is a 3-dimensional tensor [B, C, H*W], where B represents the batch size, C represents the number of channels, and H and W represent the height and width of the input, respectively.

2.2 Upsampling

Upsampling is a common method to increase the complexity of a model. U-Net(Ronneberger O et al., 2015[14]) is a end-to-end segmentation framework based on convolution backbone. We can conveniently use the U-Net architecture to train a model for lung segmentation as Fig. 2.

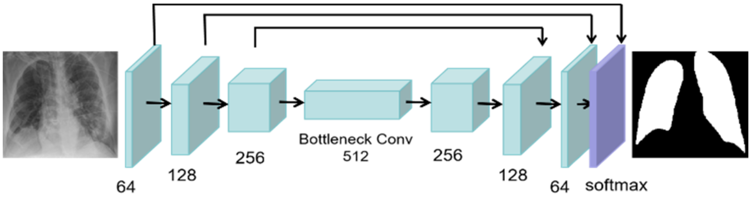

Fig. 2. U-Net model architecture used for lung segmentation

3 Proposed famework

We propose a parallel coupling and decision constraint module, as shown in Fig 3.

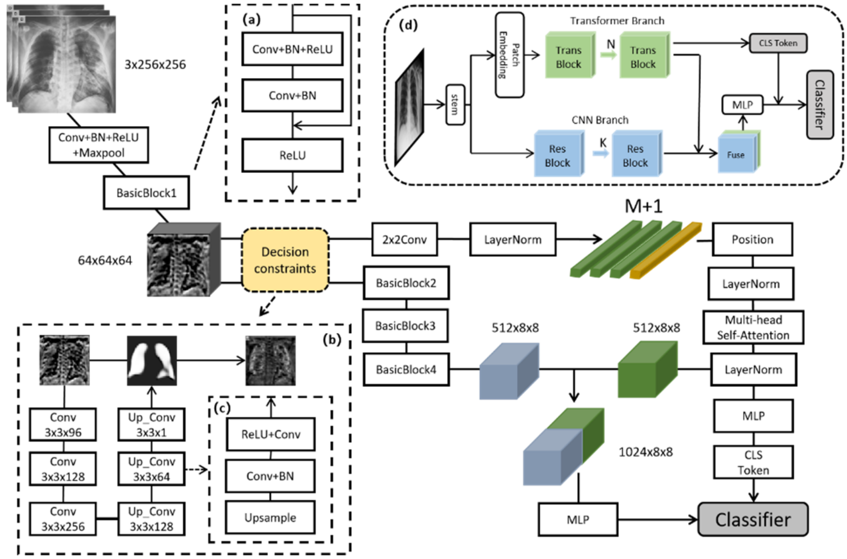

Fig. 3. The network architecture of the proposed model. (a) Residual block structure, (b) Decision constraint structure, (c) Upsampling structure, (d) Process framework

Next in this section, we will elaborate on the details of the model: decision constraints, Transformer branch structure and branch fusion details.

3.1 Decision constraints

The mentioned lung segmentation preprocessing operation is a strategy that limits the decision source, but it results in significant information loss. This paper designs a decision constraint module for the classification model and adds it to the intermediate process. The decision constraint module only consists of downsampling and upsampling, without any skip connections between shallow and deep features. The downsampling is composed of three convolutions, the upsampling is composed of linear interpolation and two convolution blocks, as Fig. 3(b). The final result will be a binary mask image generated from the input image. This binary mask image is multiplied with the input image's pixels to increase the weight of high-intensity areas and decrease the weight of low-intensity areas. Since the subsequent classification relies on this new input image, this process is adaptive in order to ensure accurate classification results. Experiments show that it has greater potential for performance improvement than direct lung segmentation pretreatment.

3.2 Transformer branch

We adopt the framework of Vit, using the Transformer Encoder as the building block structure. In this paper, Transformer Block is applied to divide 64x64 input into 2x2 patches, set depth to 2, and merge embedded features into 512x8x8 shapes, which are the same size as CNN features for feature fusion. In addition, the CLS token is reserved to predict the results together with the MLP of the CNN branch.

3.3 Proposed model Structure

In order to better integrate local features and global representation, this paper believes that the original structure of CNN and Transformer in series will be improved if it is designed as a parallel network structure, as shown in Fig. 3(d). As overall framework and architecture, the proposed model is composed of a stem module, two branches, fused features and a classifier. The stem module is used to extract the initial low-level features and then generates input images through decision constraint module sending to the two branches respectively. The CNN branch and Transformer branch are composed of repeated Residual Blocks and Transformer Blocks respectively. This concurrent structure means that CNN and transformer branches can retain local features and global representation to the maximum extent, respectively. Feature map and patch embedding are merged in the channel dimension as fusion features. Then the feature is extracted by a 2x2 convolution with a step size of 2. Finally, the fusion features are input to a classifier, while retaining the classification results of the Transformer branch CLS token. In the reasoning process, the two classifiers add the same importance weight as the output result.

4 Datasets

In this study, Dataset1 prepared by the (M. E. H. Chowdhury et al., 2020[15]) contains COVID-19, Normal, Lung opacity and Viral pneumonia. And Dataset2 prepared by (Kermany D S et al., 2018[10]), contains Normal, Bacteria pneumonia and Viral pneumonia. Besides, NIH-8 Dataset has eight findings, a total of more than 38020 images by (X. Wang et al., 2017[16]). This paper considers transfer learning on NIH-8 dataset to extract low-level CXR features to improve the general performance of Transformer model in small Datasets.

4.1 Datasets description

There are two Dataset used to verify the generalization ability of the model. The images of Dataset1, a total of 21165 CXR images. Among them, 10192 were Normal, 3616 were COVID-19 positive, 1345 were Non COVID Viral pneumonia images, and 6012 were Lung opacity CXR images. The Dataset2 contains 2572 normal, 2772 Bacterial pneumonia, and 5109 Viral pneumonia images. See Table 1 for division details.

Table 1. Details of the dataset used for training, validation, and testing.

Dataset	Datasets description			
	Types	Total	Train, Val	Test
Dataset 1	COVID-19	3616	3245	371
	Normal	10192	9108	1084

	Lung opacity	6012	5332	680
	Viral pneumonia	1345	1218	127
Dataset 2	Bacteria pneumonia	2772	2502	270
	Normal	2572	2293	279
	Viral pneumonia	5109	4612	497

5 Experimental results

As mentioned earlier, there are two different Datasets and two experiments (on plain images and mask segmented lungs images) were conducted. In addition, we trained low-level semantic on NIH-8 Dataset, activated Transformer structure.

Table 2. The classification performance of single model and combined model on dataset1,2 is compared.

Models **(Params)**	**Dataset 1** **Acc**	**Dataset 2** **Acc**
CNN-based		
ResNet18 (11M)	94.25%	89.67%
ResNet50 (23M)	94.34%	89.00%
DenseNet121 (6M)	94.21%	**91.01%**
ConvNeXt-T (27M)	92.75%	87.19%
EfficientNetV2 (20M)	**94.61%**	90.92%
Transformer-based		
ViT-T/16 (14M)	90.72%	88.53%
Pre ViT-T/16 (14M)	**91.95%**	**89.29%**
CNN-Transformer		
Res18-ViT-T/2 (26M)	93.99%	88.81%
Pre Res18-ViT-T/2 (26M)	94.25%	90.06%
Proposed (22M)	94.16%	90.06%
Pre Proposed (22M)	**95.45%**	**91.97%**

According to the performance metrics in Table 2, it can be observed that CNN networks demonstrated significant advantages in both dataset 1 and 2. Despite leveraging the pretraining weights from NIH-8, the Transformer-based model achieved an accuracy of 91.95% on dataset 1, which is still lower than the accuracy of 94.61% achieved by EfficientNetV2. Similarly, on dataset 2, the accuracy of 89.29% is lower than the accuracy of 91.01% achieved by DenseNet121. On the other hand, the serial combined model, even with the pretraining weights, can only perform on par with the CNN model. However, the Proposed model with pretraining weights achieved the highest accuracy on both datasets, reaching 95.45% and 91.97% respectively. Figure 4 also demonstrates the advantages of the parallel model with pretraining weights.

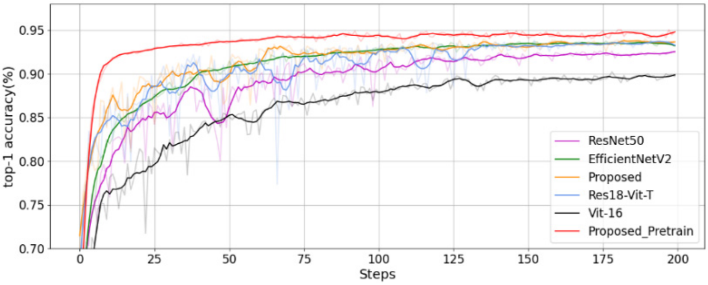

Fig. 4. Valid accuracy curve of Dataset1

According to Table 3, our proposed model performs exceptionally well in the COVID-19, Lung opacity, and Normal categories. It only lags behind the serial model in the Virus category, which could possibly be attributed to data imbalance. Overall, we believe that the parallel model has an advantage in integrating CNN and Transformer features.

Table 3. The performance of serial and parallel structures on dataset1.

Models	**Dataset 1**				
	COVID-19	**Lung opacity**	**Normal**	**Virus**	**Total**
Pre Res18-ViT-T/2	95.42%	89.71%	96.22%	**98.43%**	94.25%
Pre Proposed	**98.65%**	**92.79%**	**96.40%**	92.91%	95.45%

Table 4. The effects of lung segmentation and decision constraints on classification accuracy are compared.

Models	**Dataset1**		
	Acc	**Precision**	**F1**
Lung Segmented			
ResNet18	92.97%	93.01%	92.99%
DenseNet121	92.84%	92.90%	92.87%
EfficientNetV2	93.81%	93.81%	93.81%
Pre Res18-ViT-T/2	92.93%	92.95%	92.94%
Decision Constraint			
ResNet18	94.12%	94.19%	94.16%
DenseNet121	94.16%	94.17%	94.17%
EfficientNetV2	94.21%	94.22%	94.21%
Pre Res18-ViT-T/2	93.99%	94.03%	94.01%
Pre Proposed	**95.31%**	**95.35%**	**95.33%**

In the following experiments, we compared the effects of two approaches: directly segmenting the input images of the lungs and incorporating a decision constraint module. As the decision constraint module was designed as a plug-and-play component, we added this module to multiple different models to verify its effectiveness. As Table 4.

Compared to the data in Table 2, the lung segmentation method clearly leads to a noticeable decrease in accuracy. However, the models with the decision constraint module have not experience significant losses in accuracy. Our proposed model only dropped from 95.45% to 95.31%.

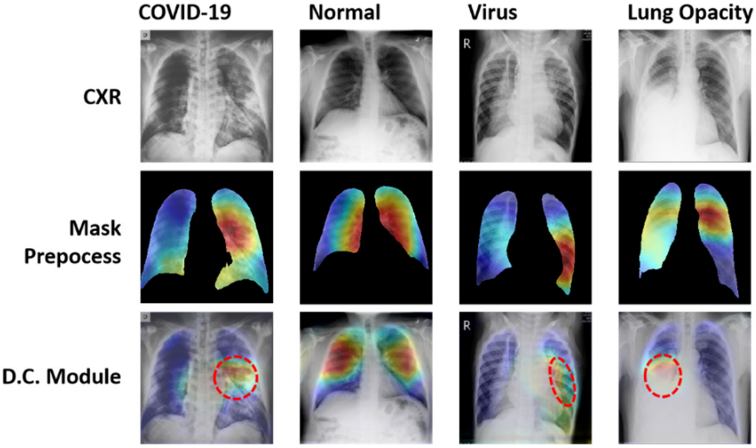

Fig. 5. Proposed model for grad cam visualization of correctly classified CXR images.

Observing the basis of network decisions is crucial. The Grad-CAM technology can generate heatmaps based on gradient information obtained from backpropagation. These heatmaps are used to validate the original chest X-ray images, lung-segmented images, and the decision constraint module through Grad-CAM heatmaps. Based on the Fig 5, the decision constraint model is capable of generating reliable visual results.

6 Conclusions

Computer-aided diagnosis is crucial in clinical medicine, and result visualization plays a key role in biomedical applications. In this study, we conducted experiments using chest X-ray images. Our parallel CNN Transformer combined model achieved a COVID-19 diagnosis accuracy of 98.65% with pre-trained weights. Additionally, our proposed decision constraint module provided a solution for heatmap visualization issues and improved upon previous research methods, striking a balance between accuracy performance and visualization performance. This technology confirms the reliability of trained models by reducing irrelevant region weights and focusing on training decisions derived from the lung region. This deep learning algorithm-based approach aids in the auxiliary diagnosis of pneumonia and alleviates the scarcity of medical resources.

Acknowledgments

This work was supported by First Class Discipline of Zhejiang-A (Zhejiang University of Finance and Economics-Statistics).

References

1. Rajpurkar P, Irvin J, Zhu K, et al. (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. doi: 10.48550/arXiv.1711.05225.
2. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3462-3471.doi: 10.1109/CVPR.2017.369.
3. Wang Wei, Hu Yiyang, Wang Xin, et al. DD-CovidNet Model for X-Ray Images Recognition of Coronavirus Disease 2019 [J] Journal of Computer-Aided Design and Computer Graphics 33(11):1649-1657. doi:10.3724/SP.J.1089.2021.18791.
4. Yang Jiezhi, Tang Wanmei, PI Jiatian, et al. Fine-Grained Pneumonia Recognition Method [J] Journal of Chongqing Normal University: Natural Science Edition, 2021, (3): 100-.doi:10.11721/cqnuj20210317.
5. Yi Sanli, Wang Tianwei, Yang Xuelian, et al. Research on ARS-CNN algorithm in the identification of COVID-19[J]. Chinese Journal of Liquid Crystals and Displays, 2021, 36(11): 1565. https://opticsjournal.net/Articles/OJ46d908a0cabf597f/References.
6. Jiang J, Lin S. (2021) COVID-19 Detection in Chest X-ray Images Using Swin-Transformer and Transformer in Transformer. https://doi.org/10.48550/arXiv.2110.08427.
7. K. S. Krishnan and K. S. Krishnan, "Vision Transformer based COVID-19 Detection using Chest X-rays," 2021 6th International Conference on Signal Processing, Computing and Control(ISPCC),Solan,India, 2021, pp. 644-648, doi: 10.1109/ISPCC53510.2021.9609375.
8. Shome, D.; Kar, T.; Mohanty, S.N.; Tiwari, P.; Muhammad, K.; AlTameem, A.; Zhang, Y.; Saudagar, A.K.J. COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare. Int. J. Environ. Res. Public Health 2021, 18, 11086. https://doi.org/10.3390/ijerph182111086.
9. Gao X, Qian Y, Gao A. (2021) COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models. https://doi.org/10.48550/arXiv.2107.01682
10. Kermany D S , Goldbaum M , Cai W , et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning[J]. Cell, 2018, 172(5):1122-1131. https://doi.org/10.1016/j.cell.2018.02.010.
11. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Abul Kashem SB, Islam MT, Al Maadeed S, Zughaier SM, Khan MS, Chowdhury MEH. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med. 2021 May;132:104319. doi: 10.1016/j.compbiomed.2021.104319. Epub 2021 Mar 11. PMID: 33799220; PMCID: PMC7946571.
12. Vaswani A , Shazeer N , Parmar N , et al. (2017) Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762.
13. Dosovitskiy A , Beyer L , Kolesnikov A , et al. (2020) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929.
14. Ronneberger O, Fischer P, Brox T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. https://doi.org/10.48550/arXiv.1505.04597.

15. M. E. H. Chowdhury et al., "Can AI Help in Screening Viral and COVID-19 Pneumonia?," in IEEE Access, vol. 8, pp. 132665-132676, 2020, doi: 10.1109/ACCESS.2020.3010287.
16. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3462-3471, doi: 10.1109/CVPR.2017.369.