Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Displays 72 (2022) 102150

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

COVID-19 CT image recognition algorithm based on transformer and CNN☆

Xiaole Fan, Xiufang Feng *, Yunyun Dong, Huichao Hou

College of Software, Taiyuan University of Technology, Taiyuan 030024, China

ARTICLE INFO

Keywords:
Bi-directional feature fusion
Transformer
CNN
COVID-19

ABSTRACT

Novel corona virus pneumonia (COVID-19) broke out in 2019, which had a great impact on the development of world economy and people's lives. As a new mainstream image processing method, deep learning network has been constructed to extract medical features from chest CT images, and has been used as a new detection method in clinical practice. However, due to the medical characteristics of COVID-19 CT images, the lesions are widely distributed and have many local features. Therefore, it is difficult to diagnose directly by using the existing deep learning model. According to the medical features of CT images in COVID-19, a parallel bi-branch model (Trans-CNN Net) based on Transformer module and Convolutional Neural Network module is proposed by making full use of the local feature extraction capability of Convolutional Neural Network and the global feature extraction advantage of Transformer. According to the principle of cross-fusion, a bi-directional feature fusion structure is designed, in which features extracted from two branches are fused bi-directionally, and the parallel structures of branches are fused by a feature fusion module, forming a model that can extract features of different scales. To verify the effect of network classification, the classification accuracy on COVIDx-CT dataset is 96.7%, which is obviously higher than that of typical CNN network (ResNet-152) (95.2%) and Transformer network (Deit-B) (75.8%). These results demonstrate accuracy is improved. This model also provides a new method for the diagnosis of COVID-19, and through the combination of deep learning and medical imaging, it promotes the development of real-time diagnosis of lung diseases caused by COVID-19 infection, which is helpful for reliable and rapid diagnosis, thus saving precious lives.

1. Introduction

COVID-19 has spread all over the world, and the diagnosis of COVID-19 by CT images of lung has been clinically verified. Relevant studies [1,2] have shown that CT images information of lung can play a vital role in the diagnosis of COVID-19. However, because of the short epidemic time in COVID-19, although there are some automatic identification methods, their schemes still have a lot of space for improvement, and they rely more on doctors' experience and consume resources. The local and overall features of the lesions are important basis for the diagnosis of COVID-19, and the diagnosis can not be made only according to the characteristics of a certain place. The analysis and diagnosis of CT images is an extremely complicated process, which requires doctors' professional knowledge and related experience. The doctor's manual experience is a time-consuming and labor-intensive process. Moreover, some CT images of COVID-19 are morphologically similar to traditional pneumonia images. The results of CT images correspond to three different types of infection: (A) novel coronavirus caused by Covid 19 virus infection, (B) common pneumonia and (C) normal control. The Fig. 1 shows the data samples.

The main characteristics of pneumonia caused by novel coronavirus in medical images are ground-glass opacities (GGO), which usually appears on both sides and around; With the progress of the disease, sometimes the paving stone sign appears, interlobular septum is thickened, and interlobular line overlaps with ground-glass opacities, which is called "crazy-paving". It is generally believed that this kind of manifestation occurs in the later stage of the disease. Others will show local hemangiectasis and other phenomena. The following picture is the main medical feature of new coronary pneumonia shown in Fig. 2.

CT images of COVID-19 not only have local features, such as local hemangiectasis and local crazy-paving, but also have global features, such as large-area ground-glass opacities. It is characterized by the integration of local and global features. At present, it is still difficult to extract the image features with relatively complex features, so it is more urgent to solve the classification problem of such medical images.

In order to solve the problem of local and global feature extraction of COVID-19 CT images, the traditional method is to extract features of different scales through multi-scale feature fusion, and the most common method is multi-scale feature pyramid network [3], which uses the receptive field of convolution kernels of different scales to learn features

☆ This paper was recommended for publication by Prof G Guangtao Zhai.

* Corresponding author at: College of Software, Taiyuan University of Technology, Taiyuan 030024, China.
E-mail addresses: f_xf_tyut@163.com (X. Feng), dongyunyun@tyut.edu.cn (Y. Dong).

https://doi.org/10.1016/j.displa.2022.102150
Received 27 October 2021; Received in revised form 25 December 2021; Accepted 4 January 2022
Available online 24 January 2022

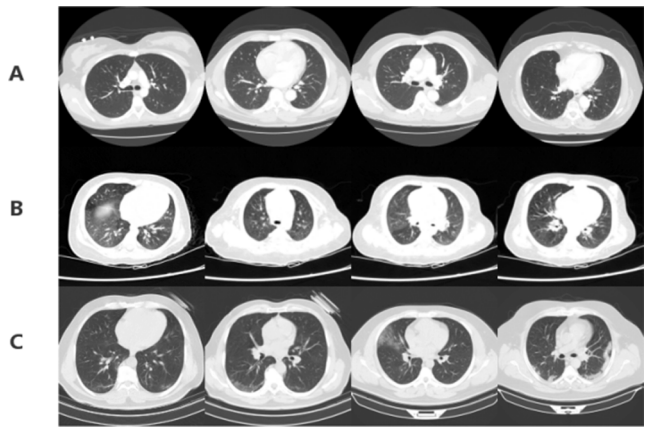
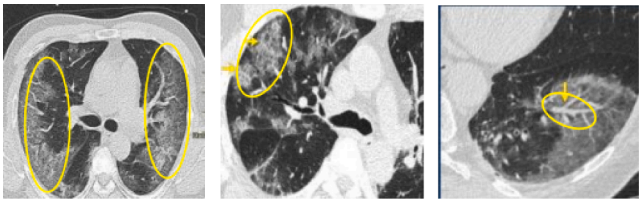


Fig. 1. Example of data set.



(a)GGO (b) crazy-paving (c) hemangiectasis

Fig. 2. Medical features of CT images in COVID-19.

of different scales. Due to the local receptive field of convolution kernels, this method is still only local features in essence. Another method is to use cascaded branch network [4]. Different branches extract features of different scales, and finally feature fusion is carried out, which is easy to understand. However, to extract the branches of global features requires a larger receptive field, either increasing the size of convolution kernel or using dilated convolution [5]. The problem is that the convolution kernel is too large, which will reduce the generalization ability of the network [6].

The image pyramid method proposed by Liu Z et al [7]. The problem of image feature extraction at different scales is solved. The images are scaled at different proportions to obtain image pyramids, and then features of different proportions are extracted from each layer of images to obtain feature maps. The pyramid of an image is a set of images arranged in pyramid shape with gradually decreasing resolution and derived from the same original image. It is obtained by down-sampling step by step, and does not stop sampling until a certain termination condition is achieved. The image is compared to pyramid layer by layer. The higher the level, the smaller the image, and the lower the resolution. Features of different scales can contain rich semantic information with high accuracy, but this overlapping pairs of image features of different scales will slow down the processing speed and increase the data volume.

The Feature Pyramid Network (FPN) proposed by Lin T Y et al [8]. A feature fusion method with different resolutions is proposed, that is, the feature map with different resolution and subsampled low-resolution features element-wise are added, so as to enhance features at different levels. Because this method only carries out cross-layer connection and element-wise addition on the basis of network, the increased computation is less and the performance is better.

Wang S H et al [9–11] have published several works on COVID-19. The Paper [9] proposed a deep rank-based average pooling network model, i. e. Deep rank-based average pooling network. The paper [10] proposed an artificial intelligence model to diagnose COVID-19 based on chest CT images. The two-dimensional fractional Fourier entropy was used to extract features and a custom deep stacked sparse autoencoder (DSSAE) model was created to serve as the classifier. Huang Z et al [12] propose a deep learning (DL) based dual-tasks network, that based on the combination of 3D CT imaging and clinical symptoms information. The above work has achieved good results.

Recently, Vision Transformer (ViT) [13] and Deit [14] have made breakthroughs in the field of computer vision, demonstrating the advantages of global processing, and have made significant performance improvement compared with CNN. However, if only the Transformer structure is used to extract features, the parameters of the network will increase greatly. When the calculation budget is limited to 1 G FLOPs, the gain of ViT will decrease. If the amount of computation increases further, CNN [15–17] and its related improvement work [18,19] will still occupy a dominant position in the field with relatively small amount of computation, which is determined by the advantages of CNN's deep convolution and point multiplication convolution in processing local features.

In order to better classify CT images in COVID-19, can we combine Transformer and CNN, make full use of their respective advantages, design an efficient network, and interact effectively with local and global specialist clinics?

Wu Haiping et al [20] introduced CNN into visual Transformer, obtained convolution vector by deep convolution, and then transformed features into query vector, value vector and key vector of Transformer. The whole structure enabled Transformer to use convolution module in the middle, and obtain 87.75% Top-1 precision on ImageNet dataset, which had fewer parameters than other transformers.

Recently, Facebook proposed LeViT network [21], which introduced convolution operation on the basis of Deit network structure, and designed the network into a structure similar to the classic LeNet architecture in CNN. First, it was a four-layer convolution, then a series of Transformer modules, and finally classification. This structure is a typical serial structure of Transformer and CNN. Some studies [22] also show that using CNN at the beginning of the network can significantly improve the network effect.

In this paper, a bi-branch network structure based on Transformer and convolutional neural network is proposed. Different from the above traditional methods, the features of different scales are not extracted by changing the size of convolution kernel, but by using the global receptive field characteristics of Transformer network, the branches of Transformer network are selected to extract the global features of images. We still use CNN's advantages in extracting local features of images, and design a bi-directional feature fusion structure, which fuses the features of Transformer branch and CNN branch, so that the network has the ability to extract more comprehensive and abundant features.

The contributions of this paper are as follows.

1) A branch network model based on CNN and Transformer structure is proposed;

2) Design a bi-directional feature fusion structure, which can fuse features extracted from two branches;

3) Innovatively apply the model proposed in this paper to COVIDx-CT—— a large-scale COVID-19 data set;

4) Carry out feature visualization experiment to verify the correctness of classification basis, thus proving the rationality of model.

2. Method

Convolutional neural networks have shown great advantages in extracting image features. It performs well in extracting image features with obvious local features, which is determined by the structure of CNN itself. CNN is a network structure composed of a convolution kernel, and its most prominent feature is the local receptive field of CNN. However, the processed image is a CT image used to diagnose COVID-19, and its medical features include obvious local lesion features and scattered overall features. Because of the particularity of the data set, this paper selects the Transformer module to extract the global features of CT images and combines them to achieve better classification results.

In this paper, a classification model based on Transformer module and Convolutional Neural Network module for extracting features of

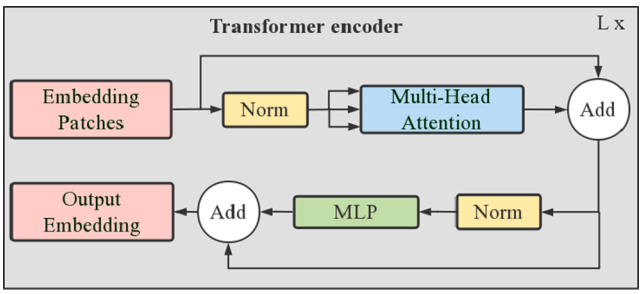


Fig. 3. Encoder model in Transformer.

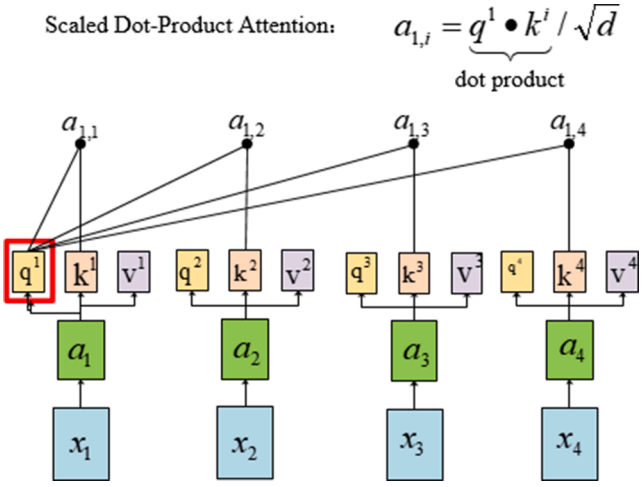


Fig. 4. Self-Attention module.

COVID-19 CT images is proposed. the specific steps are as follows.

1) the COVID-19 CT image passes through the Transformer branch module, and because of its global receptive field characteristics, the global features of the image are extracted;

2) the COVID-19 CT image passes through a convolution neural network branch module, and local features of the image are extracted by using convolution local receptive field features;

3) a bi-directional feature fusion structure is established between the two branches to fuse the features of the two branches. Here, the feature fusion is bi-directional fusion, which can extract richer and more comprehensive features and improve the classification accuracy.

4) finally, the classification vectors extracted from the two branches are fused, the loss is calculated, and the inverse gradient calculation is carried out to adjust the model parameters.

2.1. Transformer module

Transformer [23] is a model proposed by Google in 2017, which is applied in the field of natural language processing. Transformer network structure is mainly composed of attention mechanisms, and its significant feature is global receptive field. From another perspective, "Transformers" is actually a special CNN, with a global feeling field.

The basic structure of transformer is composed of two parts, including encoder and decoder. In this branch, only the encoder part of the transformer is used, and its structure is shown in Fig. 3.

Self-Attention module, the Self-Attention used in Transformer is a normalized dot product attention mechanism. As shown in Fig. 4, the feature vector of input x_i is a_i, and a_i is mapped to corresponding q_i、k_i and v_i. The Self-Attention mechanism is to perform matrix operation on each $\boldsymbol{q}$ and all $\boldsymbol{k}$, that is, dot multiplication operation. Before that, every input mapping should be normalized, so as to get the attention weight matrix.

Assuming that the input dimensions of query and key are d_k, and the value dimension is d_v (since the Transformer structure was first applied in the field of natural language processing, the symbols such as query, key and value continue to be used), then the point-multiplied operation of query and each key is calculated and divided by $\sqrt{d_k}$, and then the weight is calculated by softmax function.

$$\text{Attention}(Q, K_i, V_i) = softemax(\frac{Q^T K_i}{\sqrt{d_k}})V_i \tag{1}$$

In the field of computer vision, query, keys and values are regarded as matrices Q, K and V and the output matrix are as follows.

$$\text{Attention}(Q, K, V) = softemax(\frac{Q^T K}{\sqrt{d_k}})V \tag{2}$$

The encoder consists of two Norm and Add layers, and the calculation formula is as follows.

$$\text{LayerNorm}(X + \text{MultiHeadAttention}(X)) \tag{3}$$

$$\text{LayerNorm}(X + \text{FeedForward}(X)) \tag{4}$$

X represents the input of Multi-Head Attention or Feed Forward, and Multi-Head Attention(X) and Feed Forward(X) represent the output (The dimensions of output and input X are the same, so they can be added). The Feed Forward layer is connected by two fully connected layers. The activation function of the first layer is Relu, and the activation function of the second layer is not used. The corresponding formula are as follows.

$$\text{Max}(0, XW_1 + b_1)W_2 + b_2 \tag{5}$$

X is the input, and the dimension of the output matrix obtained by Feed Forward is consistent with X.

2.2. CNN (Convolution neural network) module

Compared with AlexNet, the improvement of VGG 19 network is that the larger convolution kernel in AlexNet is replaced by multiple convolution kernel of 3 × 3 convolution kernels, aiming at ensuring the same receptive field and improving the depth of the network. In addition, using 3 × 3 convolution kernel instead of large convolution kernel will also reduce the parameters.

The VGG 19 network has excellent performance in image classification task. The network structure is simple, and it consists of five convolution pool modules, in which convolution consists of convolution kernel of 3 × 3 and the pool layer consists of convolution kernel of 2 × 2. Because the local features extracted by this branch module need to be merged with another branch at last, for the convenience of calculation, 3 linear fully connected layers are connected after the 5 convolution pool modules, and the number of neurons in the last layer is 256 (See Table 1).

Table 1
Network parameters of improved VGG19.

Module	Convolution Kernel		Output
Input	224 × 224 × 3		
Block1	2 × Conv	3 × 3 × 64	112 × 112 × 64
	Pool	2 × 2	
Block2	2 × Conv	3 × 3 × 127	56 × 56 × 128
	Pool	2 × 2	
Block3	4 × Conv	3 × 3 × 256	28 × 28 × 256
	Pool	2 × 2	
Block4	4 × Conv	3 × 3 × 512	14 × 14 × 512
	Pool	2 × 2	
Block5	4 × Conv	3 × 3 × 512	7 × 7 × 512
	Pool	2 × 2	
FC_1	1 × 1 × 4096		4096
FC_2	1 × 1 × 1024		1024
FC_3	1 × 1 × 256		256

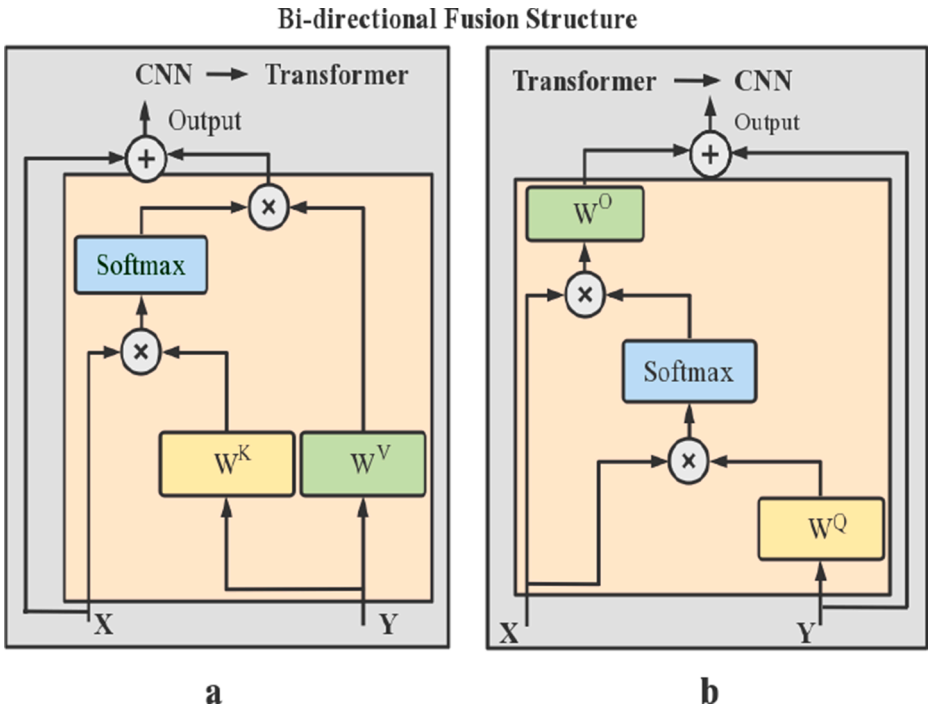


Fig. 5. Two-way feature fusion structure.

2.3. Bi-directional fusion module

The fusion layer establishes a bi-directional fusion structure between the Transformer branch and CNN branch. Refer to the bi-directional bridge structure [24], which is used to fuse global features and local features. This bi-directional fusion structure takes advantage of the advantages of Transformer and convolution CNN, and the parallel decoupling of local and global features takes advantage of the efficiency of convolution CNN in extracting local features and the global interaction ability of Transformer module. More importantly, by using a simple bi-directional fusion structure, the function of exchanging local and global information between Convolution CNN and Transformer branch is realized, which makes the efficiency and effectiveness of part and whole model well combined.

The bi-directional fusion structure is connected after the convolution pool block of Block2 and the Encoder block of the first Transformer of VGG 19 network to fuse different features. The reason why we choose to connect at these two locations is that if the connections are far apart, the parallel computing speed will be affected by the difference in computing speed between the two branches. The reason why we choose to merge in the first half of the two branches is that with the superposition of convolution layers, the receptive field gradually expands and local features gradually lost, while the global features collected in the first half of Transformer are more comprehensive. With the continuous operation of attention mechanism, the global features will be lost along with layer upon layer superposition.

In CNN → Transformer, parallelize CNN and Transformer, and connect them through self-attention structure (show in Fig. 5). CNN takes an image $X \in R^{H\times W\times 3}$ as input and extracts local features. Transformer takes learnable parameters as input, which is expressed as $Y \in R^{M\times d}$, where d and M are the dimensions and number of input vectors, respectively. To fuse with CNN, d and M are selected to have the same dimensions as the target fusion layer, and the input original image is convolved to obtain the initial Y_0.

CNN and Transformer are connected by a bi-directional feature fusion structure, in which local and global features are fused bi-directionally. Next, we will discuss the concrete structure of the bi-directional feature fusion structure.

As shown in the left figure of Fig. 5, it is a unidirectional structure of CNN → Transformer, where attention mechanism is used to fuse local features (from CNN) and global features (from Transformer). Feature vector fusion is carried out in CNN layer with few channels.

The local feature graph is denoted as X, and the global label is denoted as *Y*. They are divided into $X \in [X_h]$ and $Y \in [Y_h]$ $(1 \le h \le H)$, and *h* refers to the number of heads of multi-head self-attention. Local to global fusion is defined as follows.

$$\mathrm{head_i} = \mathrm{Attention}(Y_h W_h^Q, X_h, X_h) \tag{6}$$

$$Y^{out} = Y + [\mathrm{Concat}(\mathrm{head_1}, \cdots, \mathrm{head_h})]\mathrm{W^o} \tag{7}$$

W_h^Q is the projection matrix of query in multiple headers, W^o is used to combine multiple headers, and $\mathrm{Attention}(Q, K, V)$ is the standard Attention function on *Q*, *K* and *V*, as shown in formula (2).

Global input feature *Y* is *Q* and local input feature *X* is *K* and *V*. W_h^Q and W^O are applied to the global feature Y. Similarly, the formula for calculating the feature fusion structure from global to local is as formula (8) and (9).

$$\mathrm{head_i} = \mathrm{Attention}(X_h, Y_h W_h^K, Y_h W_h^V) \tag{8}$$

$$\mathrm{X}^{out} = \mathrm{X} + [\mathrm{Concat}(\mathrm{head_1}, \cdots, \mathrm{head_h})] \tag{9}$$

In which W_h^K and W_h^V are projection matrices of keys and values. The local feature *X* is query and the global feature *Y* is key and a value. The schematic diagram of feature fusion in this direction is shown in the right figure of Fig. 5 (Transformer → CNN).

Input and output: CNN-Transformer block has two inputs: (a) local feature graph $X \in R^{hw\times C}$, which has *C* channels and *hw* spatial positions ($hw = h \times w$ where *h* and *w* are the height and width of feature graph) and (b) global mark $Y \in R^{M\times d}$, where *M* and *d* were the number and dimension of feature blocks, respectively.

2.4. The whole model design and network structure

In this paper, the traditional serial structure is changed into a parallel structure, and a parallel mode combining Transformer and CNN is

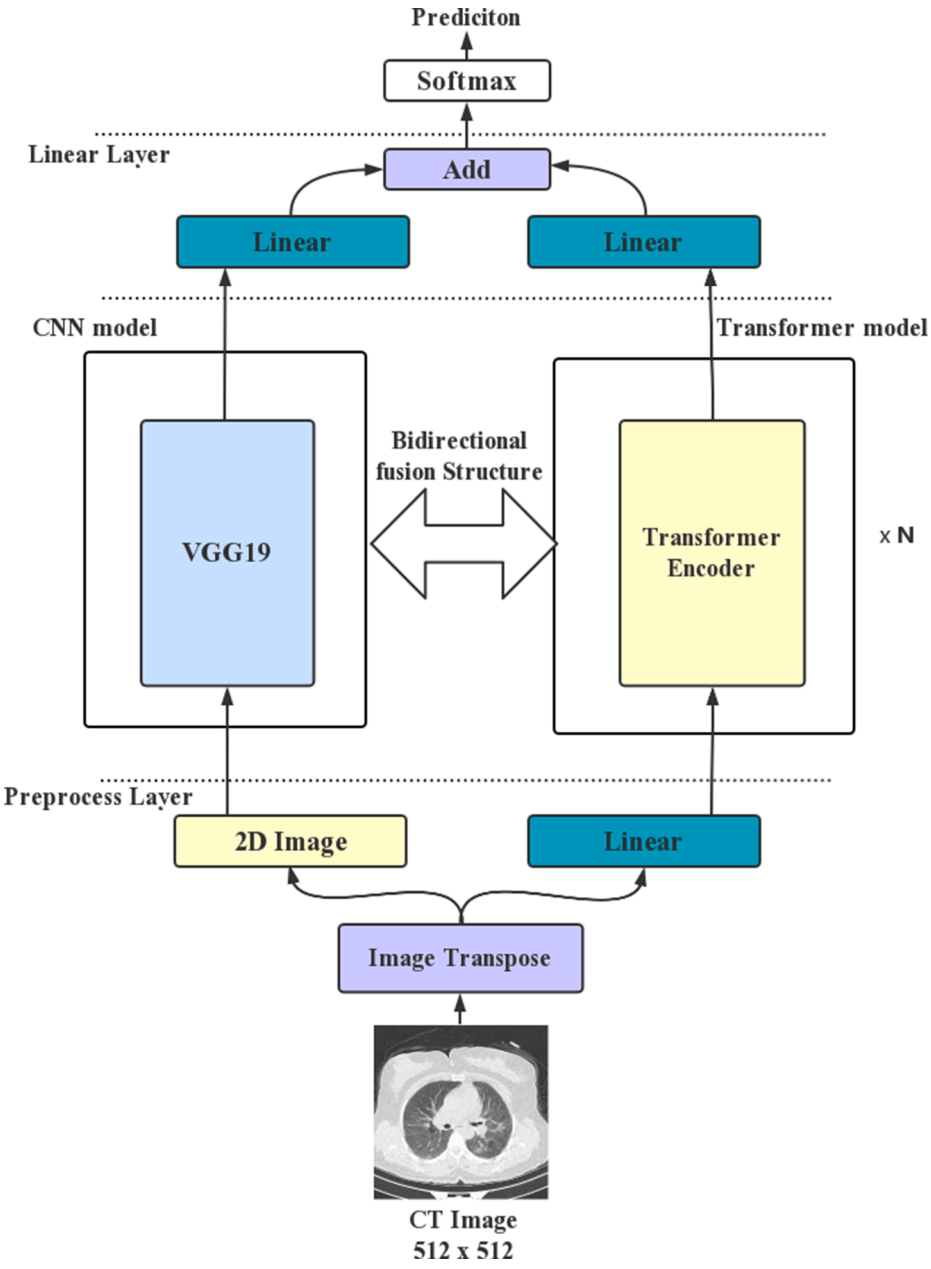


Fig. 6. Overall structure of the model.

proposed. As shown in Fig. 6, the input of Transformer is the vector of the whole picture, and the receptive field is the global receptive field, so it is possible to take advantage of the advantages of Transformer in extracting global features. At the same time, CNN can extract local features well, so it constructs a bi-branch parallel network structure, and realizes the parallelization of the roll integration branch and the Transformer branch.

In the middle of the bi-branch parallel structure, a bi-directional feature fusion structure is designed, as shown in Fig. 6, which utilizes the principle of cross fusion to achieve the effect of different feature fusion.

3. Experiment and result analysis

3.1. Experimental environment and experimental setup

In this paper, a 64-bit operating system Ubuntu-18.04.1 is adopted to implement the algorithm. The 3-card parallel training was conducted on Intel(R) Xeon(R) CPU E 5–2695 and NVIDIA TITAN V high-performance GPU, and each graphics card was executed on a computer with a storage capacity of 12 GB and 16 GB memory. Under Pytorch 1.7.0, CUDA 11.0 and CUDNN 7.6, the model was built and trained. Pre-training was conducted on ImageNet dataset, and the initial learning rate was set to 0.0005. The optimization algorithm uses the random momentum gradient descent algorithm, and batch setting is 128. The model converged at 48th epoch.

3.2. Dataset

In this study, COVIDx-CT is introduced, which is a benchmark CT image data set, derived from CT imaging data collected by China National Bioinformatics Center, including 194,922 images of 3745 patients aged 0 to 93 years (median age 51 years), which has been strongly clinically verified. In COVIDx-CT benchmark data set, the results of chest CT volume correspond to three different infection types: novel coronavirus caused by Covid 19 virus infection, common pneumonia and normal control. The distribution of patients with three infection

Table 2
Data set segmentation.

Type	Normal	Pneumonia	COVID-19	Total
train	35,996	25,496	82,286	143,778
val	11,842	7400	6244	25,486
test	12,245	7395	6018	25,658

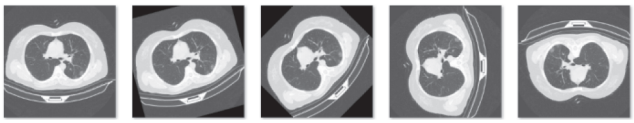


Fig. 7. Example diagram of data set preprocessing.

types in training, verification and testing is shown in Table 2.

By means of data enhancement, the sample size of the data set is expanded, and the data is expanded by rotating at different angles (15°, 45°, 90° and 180°). The enhanced data is shown in Fig. 7 and normalized. The reason for expanding the data set is that training the model of Transformer structure needs a large amount of data to achieve better results.

3.3. Experimental result

The process of training loss and precision convergence is as shown in Figs. 8-9. Because the model has been pre-trained in ImageNet data set in advance, it has good feature extraction ability and fast convergence speed during the training process.

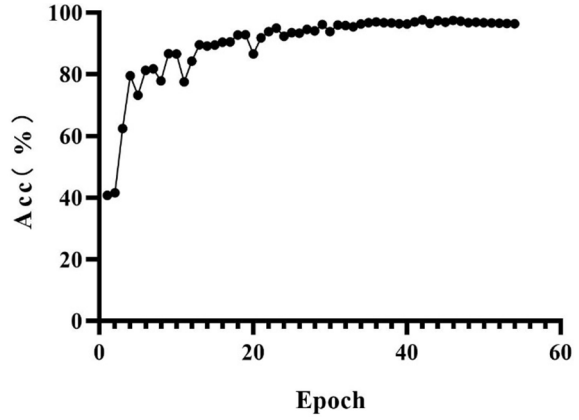


Fig. 8. Accuracy of training network.

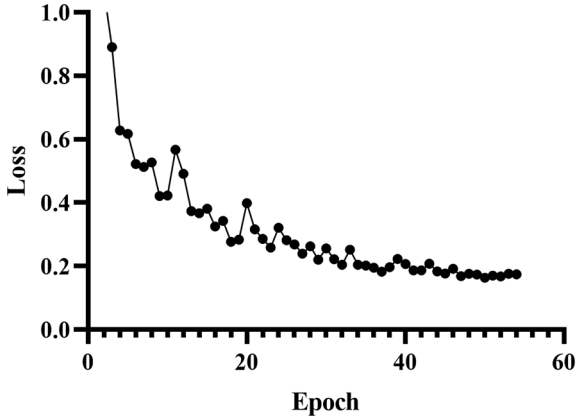


Fig. 9. Loss of training network.

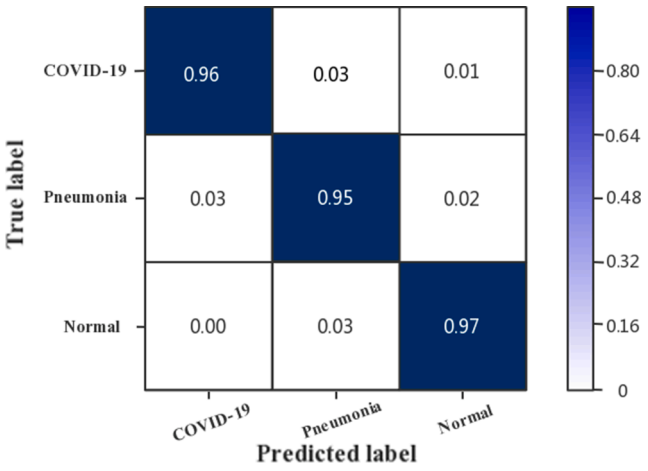


Fig. 10. Confusion matrix obtained by Trans-CNN Net test.

As shown in Fig. 10, the confusion matrix of the model shows that after loading the trained model, the classification accuracy of the model for common pneumonia is about 95%, and the recognition accuracy for COVID-19 is about 96%, and the overall accuracy is about 97%.

3.4. Experimental evaluation index

When measuring the performance of our proposed models, we used five indicators in our experiment, including specificity, sensitivity, accuracy, precision and F 1 score. True positive (TP), false positive (FP), true negative (TN) and false negative (FN) are four main components involved in the calculation of the above indexes. TP indicates the number of correctly classified pneumonia caused by COVID-19; TN represents the number of normal images correctly classified; FP indicates the number of images that are misclassified as pneumonia caused by COVID-19; FN indicates the number of images that are misclassified as normal.

Specificity reflects the performance of our models in identifying normal images from the test set, and can be passed.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity is the percentage of true positive (COVID-19 confirmed pneumonia) images that are correctly recognized and can be defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Accuracy proves our model's ability to classify correctly, which can be written as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Precision describes the percentage of predicted patients are true patients by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F1 score measures the classification ability:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Compared with the traditional Resnet-152 network, which has better classification effect on COVIDx-CT, the classification accuracy of pure Transformer network Deit on this data set is 95.2% and 75.8%, respectively, and the classification accuracy of this network is better (See Table 3).

Table 3
Comparison between Tran-CNN Net and other classification models.

Model Construction	Params (M)	FLOPs	Acc (%)
ResNet-152	60.2	11.8 M	95.2%
Deit-B	86.1	291.3	75.8%
Tran-CNN Net	92.6	301 M	96.7%

Table 4
Comparison with proposed methods.

Model	Specificity	Recall	F1	Precision	Accuracy
K-ELM [25]	0.5682	0.6136	0.5952	0.5909	0.5814
Wang [26]	0.6700	0.7400	–	–	0.7310
Li [27]	0.9000	0.9600	–	–	0.9600
Mangal [28]	–	1.0000	–	–	0.9050
ResGNet-C [29]	0.9591	0.9733	**0.9665**	0.9621	0.9662
CovidCTNet [30]	–	0.8300	–	–	0.9000
ViT-B [31]	09,320	0.9380	0.9460	0.9530	0.9760
X Wang [32]	0.9163	0.9323	0.9050	0.9401	0.9241
Zhang J [33]	0.8935	0.9012	0.9228	0.8870	0.9210
Trans-CNN Net (Ours)	**0.9601**	**0.9776**	0.9636	**0.9745**	**0.9673**

In order to further verify our proposed method, the proposed method is compared with some existing methods. Results As shown in Table 4, the model proposed in this paper performs best in Specificity, Recall, F1, Precision and Accuracy.

3.5. Visualization of classification basis

In order to better verify the validity of the model, Grad-weighted class activation mapping (Grad-CAM) is used to visualize the key feature areas of model classification, and the results are shown in Fig. 11.

4. Discussion

In this study, we designed and implemented a bi-branch feature fusion framework based on CNN and Transformer to identify CT images. The experimental results show that this model is an effective model for COVID-19 detection using CT images. The main advantage of this method is that it makes full use of the advantages of Transformer and CNN in extracting features of different scales, and builds a bi -branch network.

The result is better than single CNN structure network and single Transformer network. When designing the bi-branch network, our experiment compares the combination results of different CNN and Transformer networks, and finds that the combination of traditional Transformer network and VGG 19 is the best as Fig. 12.

Global features refer to the overall attributes of an image. Common global features include color features, texture features, shape features, etc. In order to explore the decreasing effect of Transformer in extracting global features, we designed experiments to obtain activation map of Transformer at different stages. As shown in Fig. 13, column a is the original image, and column b is the feature map output by Block1 of Transformer. We can clearly see that Block1 of Transformer has abundant global features. Column c is the feature map output by Block3 of Transformer, and the global features are greatly reduced by the self-attention mechanism of Transformer; Column d is the output characteristic diagram of network Block 5. We can see that after passing through the last block of the transformer, it shows more concentrated local features. As we expected, with the increasing number of Transformer layers, the global characteristics gradually decrease.

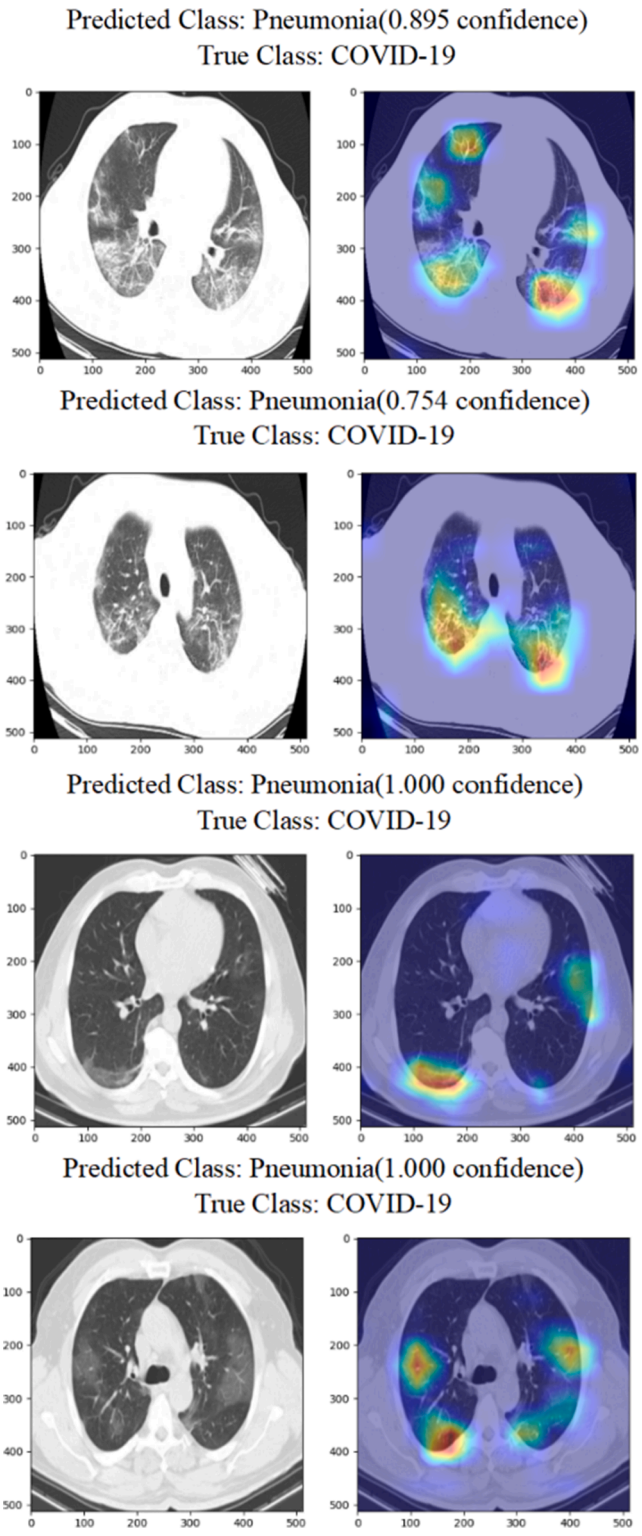


Fig. 11. Comparison diagram of feature visualization.

At the same time, a bi-directional feature fusion structure is proposed. As we know, with the increase of network layers, the information of the original data will be gradually lost, while the bi-branch network can extract different features, and fuse the features of different branches through the bi-directional feature fusion structure to reduce the loss of useful feature information. When the design features are fused in two branch nodes, we have repeatedly experimented with multiple fused

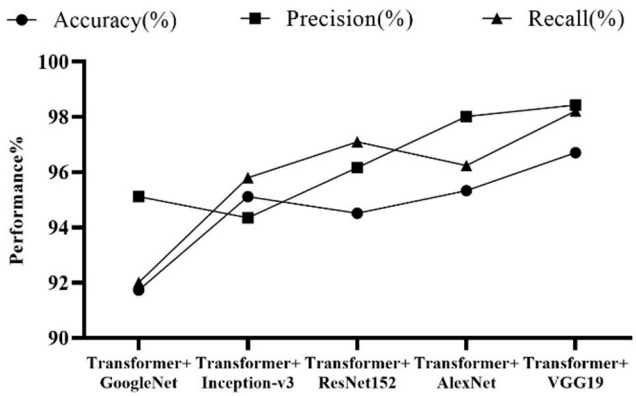


Fig. 12. Comparison of the combination of Transformer and different CNN structures.

nodes, and the results show that the fusion effect is the best after the first encoder of Transformer and the second Block of CNN as the Table 5.

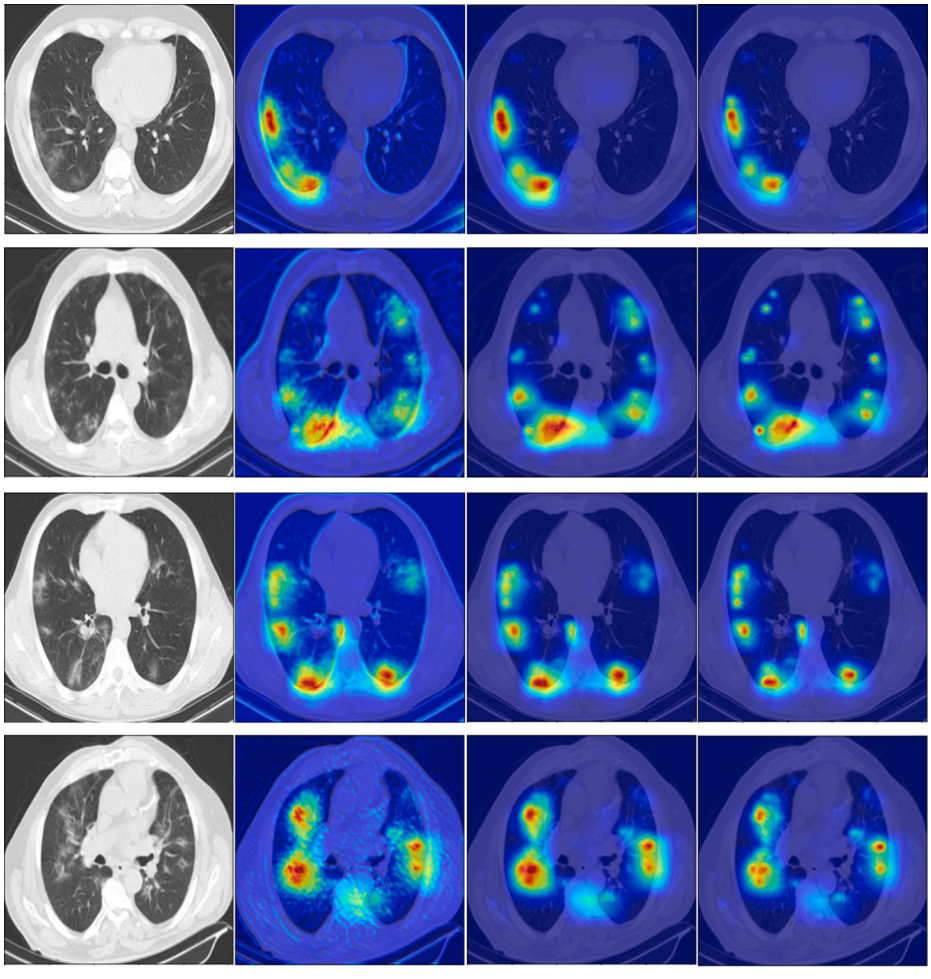
5. Conclusions

In this paper, a bi-branch feature fusion network structure based on Transformer module and Convolutional Neural Network module is proposed. Using the advantages of local features and global features of the two branches respectively, the features of CT images are extracted, and the features of the two branches are fused bi-directionally in the fusion layer, so that the network can process data in parallel, improve the running speed of the network, and achieve better classification results. The method is simple in structure and strong in generalization ability, and can extract the features with local and global features. It performs well. In the task of COVID-19 classification using CT images, and the classification accuracy is 96.7%. To sum up, this research is of great significance to the classification of medical image.

However, in this experiment, when classifying CT images, it is based on one of multiple CT images of patients, which has fewer features and has the problem of incomplete information of patient diagnosis results. Therefore, on the basis of perfecting the network, we can consider three-dimensional reconstruction on multiple CT images of the same patient, and further explore this in future work.

Table 5
Performance metrics of different layer connections.

Possible combinations	Performance metrics (%)				
	Trans-Block1	Trans-Block2	Trans-Block3	Trans-Block4	Trans-Block5
CNN-Block1	96.50	95.98	94.03	95.11	93.65
CNN-Block2	**96.71**	96.01	94.32	95.71	92.23
CNN-Block3	96.25	94.36	92.71	93.28	90.27
CNN-Block4	95.98	93.21	92.88	90.45	90.36
CNN-Block5	95.01	93.11	91.42	89.33	87.21



a b c d

Fig. 13. Transformer extracts that effect of global feature.

CRediT authorship contribution statement

Xiaole Fan: Conceptualization, Data curation, Methodology, Software, Writing – original draft. **Xiufang Feng:** Supervision, Validation, Writing – review & editing, Funding acquisition. **Yunyun Dong:** Writing – review & editing. **Huichao Hou:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank all the anonymous reviewers for their valuable comments and constructive suggestions. This work is supported by Shanxi Provincial Key Research and Development Project (201903D121121).

References

[1] Z. Ye, Y. Zhang, Y.i. Wang, Z. Huang, B. Song, Chest CT manifestations of new coronavirus disease 2019 (COVID-19): a pictorial review[J], Eur. Radiol. 30 (8) (2020) 4381–4389.
[2] A.J. Rodriguez-Morales, J.A. Cardona-Ospina, E. Gutiérrez-Ocampo, R. Villamizar-Peña, Y. Holguin-Rivera, J.P. Escalera-Antezana, L.E. Alvarado-Arnez, D.K. Bonilla-Aldana, C. Franco-Paredes, A.F. Henao-Martinez, A. Paniz-Mondolfi, G.J. Lagos-Grisales, E. Ramírez-Vallejo, J.A. Suárez, L.I. Zambrano, W.E. Villamil-Gómez, G. J. Balbin-Ramon, A.A. Rabaan, H. Harapan, K. Dhama, H. Nishiura, H. Kataoka, T. Ahmad, R. Sah, Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis[J], Travel Med. Infect. Dis. 34 (2020) 101623, https://doi.org/10.1016/j.tmaid.2020.101623.
[3] Z. Lin, Z. Luo, L. Zhao, et al., Multi-scale convolution target detection algorithm with feature pyramid[J], J. ZheJiang Univ. (Eng. Sci.) 53 (3) (2019) 533–540.
[4] Cheng Weiyue, Zhang Xueqin, Lin Kezheng, et al. Deep Convolutional Neural Network Algorithm with Fusing Global and Local Features. [J/OL]. Journal of Frontiers of Computer Science and Technology: 1-11 [2021-09-02].
[5] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.071 22, 2015.
[6] H.T. Cheng, L. Koc, J. Harmsen, et al., Wide & deep learning for recommender systems[C], in: Proceedings of the 1st workshop on deep learning for recommender systems, 2016, pp. 7-10.
[7] B. Singh, L.S. Davis, An analysis of scale invariance in object detection snip[C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018:) 3578–3587.
[8] T.Y. Lin, P. Dollár, R. Girshick, et al., Feature pyramid networks for object detection[C], Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017:) 2117–2125.
[9] S.-H. Wang, M. Attique Khan, V. Govindaraj, S. L. Fernandes, Z. Zhu, Y.-D. Zhang, Deep rank-based average pooling network for COVID-19 recognition[J], Comput. Mater. Continua 70 (2) (2022) 2797–2813.
[10] S.H. Wang, X. Zhang, Y.D. Zhang, DSSAE: Deep stacked sparse autoencoder analytical model for COVID-19 diagnosis by fractional Fourier entropy[J], ACM Trans. Manage. Inform. Syst. (TMIS) 13 (1) (2021) 1–20.
[11] S.-H. Wang, Z. Zhu, Y.-D. Zhang, PatchShuffle convolutional neural network for COVID-19 explainable diagnosis[J], Front. Public Health 9 (2021).
[12] Z. Huang, X. Liu, R. Wang, M. Zhang, X. Zeng, J. Liu, Y. Yang, X. Liu, H. Zheng, D. Liang, Z. Hu, FaNet: fast assessment network for the novel coronavirus (COVID-19) pneumonia based on 3D CT imaging and clinical symptoms[J], Appl. Intell. 51 (5) (2021) 2838–2849.
[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv preprint arXiv:2010. 11929, 2020.
[14] H. Touvron, M. Cord, M. Douze, et al., Training data-efficient image transformers & distillation through attention[C], In: International Conference on Machine Learning. PMLR, 2021, 10347-10357.
[15] A. Howard, M. Sandler, G. Chu, et al., Searching for mob-ilenetv3[C], Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 1314–1324.
[16] A.G. Howard, Z.u. Menglong, C. Bo, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv, 2017.
[17] M. Sandler, A. Howard, Zhu Menglong, et al., Mobilenetv2: Inverted residuals and linear bottlenecks [C], In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:4510-4520.
[18] K. Han, Y. Wang, Q. Tian, et al., Ghostnet: More features from cheap operations[C], Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 1580–1589.
[19] Li Yunsheng, Chen Yinpeng, Dai Xiyang, et al. MicroNet: Improving Image Recognition with Extremely Low FLOPs[J]. arXiv preprint arXiv:2108. 05894, 2021.
[20] Wu Haiping, Xiao Bin, Codella N, et al. Cvt: Introducing convolutions to vision transformers[J]. arXiv preprint arXiv:2103.15808, 2021.
[21] B. Graham, A. El-Nouby, H. Touvron, et al., LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference[J]. arXiv preprint arXiv:2104.01136, 2021.
[22] T. Xiao, M. Singh, E. Mintun, et al., Early convolutions help transformers see better [J]. arXiv preprint arXiv:2106.14881, 2021.
[23] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need[C], Advances in neural information processing systems, 2017, p. 5998-6008.
[24] Y. Chen, X. Dai, D. Chen, et al., Mobile'-Former: Bridging MobileNet and Transformer[J]. arXiv preprint arXiv: 2108. 05895, 20.
[25] S. Lu, Z. Lu, J. Yang, M. Yang, S. Wang, A pathological brain detection system based on kernel based ELM[J], Multimedia Tools Appl. 77 (3) (2018) 3715–3728.
[26] S. Wang, B.o. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, B.o. Xu, A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19) [J], Eur. Radiol. 31 (8) (2021) 6096–6104.
[27] L. Li, L. Qin, Z. Xu, et al., Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT[J], Radiology (2020).
[28] Mangal A, Kalia S, Rajgopal H, et al. CovidAID: CO'VID-19 detection using chest X-ray[J]. arXiv preprint arXiv:2004.09803, 2020.
[29] X. Yu, S. Lu, L. Guo, S.-H. Wang, Y.-D. Zhang, ResGNet-C: A graph convolutional neural network for detection of COVID-19[J], Neurocomputing 452 (2021) 592–605.
[30] T. Javaheri, M. Homayounfar, Z. Amoozgar, et al., Covidctnet: An open-source deep learning approach to identify covid-19 using ct image[J]. arXiv preprint arXiv:2005.03059, 2020.
[31] K.S. Krishnan, K.S. Krishnan, Vision Transformer based COVID-19 Detection using Chest X-rays[C], in: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). IEEE, 2021: 644-648.
[32] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, C. Zheng, A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT[J], IEEE Trans. Med. Imaging 39 (8) (2020) 2615–2625.
[33] J. Zhang, Y. Chu, N. Zhao, Supervised framework for COVID-19 classification and lesion localization from chest CT[J], Ethiopian J. Health Dev. 34 (4) (2020).