Hindawi
International Journal of Antennas and Propagation
Volume 2022, Article ID 5072666, 8 pages
https://doi.org/10.1155/2022/5072666

Research Article

Identification of Pneumonia in Chest X-Ray Image Based on Transformer

Yongjun Ma [1,2] **and Wei Lv** [2]

[1]*City University of Macau, Macau 999078, China*
[2]*Zhuhai College of Science and Technology, Zhuhai 519040, China*

Correspondence should be addressed to Wei Lv; luwei@zcst.edu.cn

Received 22 May 2022; Accepted 4 July 2022; Published 1 August 2022

Academic Editor: Xianpeng Wang

The research of application models based on traditional convolutional neural networks has gradually entered the bottleneck period of performance improvement, and the improvement of chest X-ray image models has gradually become a difficult problem in the study. In this paper, the Swin Transformer is introduced into the application model of pneumonia recognition in chest X-ray images, and it is optimized according to the characteristics of chest X-ray images. The experimental results based on the model in this paper are compared with those of the model built with the traditional convolutional neural network as the backbone network, and the accuracy of the model is proved to be greatly improved. After the comparison experiments on two different datasets, the experimental results show that the accuracy of the model in this paper improves from 76.3% to 87.3% and from 92.8% to 97.2%, respectively. The experiments show that the accuracy of image enhancement based on the features of chest X-ray images in this model will be higher than the accuracy without image enhancement. In the experiments of this paper, the identification decision factors in the chest X-ray images were extracted by grad-cam combined with a transformer to find the corresponding approximate lesion regions.

1. Introduction

Pneumonia is a common and dangerous disease that is mainly caused by viruses, bacteria, or fungi. If left untreated, its mortality rate is high. According to the literature, pneumonia is one of the ten deadliest diseases in the United States and has a higher mortality rate in developing countries [1]. Chest X-ray imaging (hereafter referred to as CXR) is widely used in general routine examinations because it is not only low cost, but also its radiation is less harmful than computed tomography. Relevant papers indicate that the mean effective radiation dose per exam of CXR is about 0.04 ± 0.19 msv, while the principle of computed tomography is that X-rays penetrate the human body for multiple times for tomography, so the mean effective radiation dose per exam can reach 1.09 ± 1.11 msv, about 25 times that of CXR [2]. Doctors often use CXR as an important aid in diagnosing pneumonia. In today's world, artificial intelligence is playing a huge role in the transformation of science, industry, and society, and its techniques are widely used in medical image processing. The application and improvement of artificial intelligence in CXR to identify pneumonia can assist doctors in making the correct diagnosis, help them speed up the diagnosis, reduce the proportion of missed and misdiagnosis, and be of great importance in saving lives.

Since the explosive development of deep learning in 2012, amazing achievements have been made in the research and application of artificial intelligence. Compared with other machine learning algorithms, deep learning algorithms can rely on their own learning methods for feature extraction. Deep learning has achieved great success in many fields such as computer vision, natural language processing, and big data analysis. In addition, it has become a mainstream approach to machine learning and has achieved record-breaking results in various competitions in artificial

intelligence. Deep learning can be traced back to AlexNet in 2012 [3]. The accuracy of this convolutional neural network algorithm, which won the championship in the famous international image classification competition ImageNet, has been improved by more than ten points compared with other algorithms in the past. It uses many methods for the first time, uses ReLu as a nonlinear activation function, uses dropout to prevent overfitting, uses data enhancement, and so on. After AlexNet, there have been many excellent convolutional neural networks. VGGNet is a convolutional neural network developed by the Visual Geometry Group of Oxford University on the basis of AlexNet [4]. The improvement of VGGNet is that it uses a smaller convolution kernel and a deeper network structure, which enhances the feature learning ability of the convolutional neural network, which also verifies the advantages of small convolution kernels and can improve network performance by deepening the network structure. In addition, VGGNet uses the multi-Scale method to train and predict, reducing the occurrence of model overfitting and improving the prediction accuracy. Inspired by the Network in Network theory, the concept of the Inception module emerged, that is, a convolutional layer contains multiple convolutional operations of different sizes. A typical convolutional neural network with Inception is GoogLeNet [5]. In addition, two auxiliary classifiers are added to the middle layer of GoogLeNet to strengthen supervision information and alleviate the problem of gradient disappearance. In simple theory, the deeper the network level, the more complex feature extraction can be carried out, so better results should be obtained. But in fact, it was found in the experiment that there was a problem of degradation after the network was deepened to a certain extent, that is, after a large increase in the network depth, the accuracy began to saturate and degrade. The main reason is that when the data are transmitted in a deep network, the gradient becomes smaller and gradually disappears, making it impossible to perform the backpropagation algorithm, so it is difficult for the network to train and find a good parameter after deepening the level to a certain extent. For this reason, He et al. proposed a residual unit with a "short-circuit connection" structure to solve this degradation problem, instead of directly connecting each layer. ResNet is modified on the basis of VGGNet, and it uses residual units [6]. Compared with VGGNet, it adds a "short-circuit connection" mechanism between every two layers, which gives an implementation idea for building a much deeper network. In addition to the ways of deepening the network such as ResNet and widening the network such as GoogLeNet to improve the effect, there are also multiplexing schemes, the typical representative is DenseNet, which can achieve better results while achieving fewer parameters [7]. Other scholars have proposed EfficientNet, which is based on an artificial neural network to obtain the optimal composite coefficient of network depth, network width, and image resolution [8].

With the research and development of the convolutional neural backbone network, it has also promoted the improvement of medical image processing model capabilities. As early as 2017, Wang's team built medical image processing models based on the classic convolutional neural network AlexNet [3], VGGNet [4], GoogLeNet [5], and ResNet [6] in deep learning, and tested and compared them on the public CXR dataset named Chest X-ray. Through their research and experiments, it was proved that resnet50 has the best effect of disease identification in CXR compared with other backbone networks [9]. Yao et al. optimized the convolutional neural network DenseNet [7], and the model they proposed was tested on the Chest X-ray dataset and achieved ideal results [10]. Later, Rajpurkar and other scholars built a 121-layer network based on the convolutional neural network DenseNet and used the weighted cross entropy as the loss function to propose the chexnet model for medical image classification. The model was tested with a higher accuracy score than four human medical imaging experts correctly judged [11]. Later, many scholars further improved the models based on the convolutional neural network according to the features of CXR [12–20]. But accuracy of models began to encounter bottlenecks, and there are still some unsolved or imperfect problems in the current models.

In this paper, a new model scheme based on the backbone network of the new transformer and optimized according to the features of CXR will be proposed, and it can greatly improve the accuracy of identification of pneumonia in CXR. The image enhancement and parameter optimization scheme are designed based on the features of CXR, and the lesion area is found to the greatest extent from the decision factors of transform. Experiments in this paper show that under the same circumstances, the model for identification of pneumonia in CXR based on the transformer backbone network has higher accuracy than that based on the traditional convolutional neural backbone network. The image enhancement scheme for CXR in this model will play a positive role in improving the accuracy rate of the model.

Through the research in this paper, the bottleneck problem of improving the accuracy of the model for identification of pneumonia in CXR based on the traditional convolutional neural network can be overcome, and better results can be achieved. To sum up, the research in this paper has its value both theoretically and practically in the identification of pneumonia and even more diseases in CXR.

2. Proposed Scheme

In order to better compare the difference between the model for identification of pneumonia in CXR based on transformer backbone network and the models based on traditional convolutional neural backbone network, the experiment in this paper was done on the Chest X-ray data set [9] and CXR images (pneumonia) data set [21], because a large number of scholars used these data sets when testing the models based on the traditional convolutional neural backbone network. It should be noted that the former data set comes from the National Institutes of health, and the latter data set comes from Guangzhou Women and Children's Medical Center, and these data sets are publicly available for free use in scientific research.

The chest X-ray data set is a data set of more than 100000 anonymous chest X-ray images released by the National Institutes of health to the scientific community. The copyright of this data set is announced on https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community, "The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease." The number of samples in the Chest X-ray data set [9] is shown in Table 1.

In the CXR images (pneumonia) data set [21], there are 5,856 anonymous chest X-ray images from Guangzhou Women and Children's Medical Center with "license CC BY 4.0". The text of the CC BY 4.0 was retrieved from https://creativecommons.org/licenses/by/4.0/, and for more information, view the full license text at https://creativecommons.org/licenses/by/4.0/legalcode. This data set is divided into two categories: pneumonia and normal. The number of samples in the CXR images (pneumonia) data set is shown in Table 2.

The model used in the experiment in this paper is based on the Swin Transformer backbone network [22] and optimizes the CXR accordingly. The basic steps are as follows (Figure 1): in addition to the obvious feature of a gray-scale image, CXR generally has its own characteristics such as low brightness, poor contrast, and high noise, so the first step is to improve the brightness, contrast, and suppress noise of the image according to the features of CXR. The second step is to obtain the best parameters of the model, the images are divided into a training set and validation set, normalize the images in the training set, and after random scaling, clipping, and flipping send them to the transformer network and fully connected network for training to obtain the best parameters of this model. The purpose of normalizing images is to facilitate the speedy contingency of the network. The purpose of random scaling, clipping, and flipping is to make the model not "see" the same image twice during training, so it has better generalization ability. A transformer network is used for feature extraction and a fully connected network is used for classification. In the third step, the images in the validation set are scaled and sent to the transformer network with trained parameters for feature extraction, and then send to the fully connected network with trained parameters for classification. The fourth step is to extract the decision factor from the Transformer network. The last step is to map the decision factor to the original image to output the lesion area.

The first was to do experiments with the model in this paper on the Chest X-ray data set, and then the experimental results are compared with the experimental results of models based on AlexNet [3], GoogLeNet [4], VGGNet16 [5], and ResNet50 [6] from the Wang's team on the same data set, the experimental results of model based on DenseNet [8] from Yao and other scholars, and the experimental results of model based on DenseNet121 [8] from Rajpurkar et al.

In order to verify the effectiveness of image enhancement according to the features of CXR, a comparison experiment between enhanced and nonenhanced images in the preprocessing with the model based on the Transformer backbone network was carried out, and the two experimental results were compared. As shown in Figure 2, it is a contrast map for CXR enhancement, in which the left side is before enhancement, and the right side is after enhancement.

Table 1: The number of samples in Chest X-ray data set.

Focus of infection	Samples
Atelectasis	5789
Cardiomegaly	1010
Effusion	6331
Infiltration	10317
Mass	6046
Nodule	1971
Pneumonia	1062
Pneumothorax	2793
Normal	84312

Data set source: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community.

Table 2: The number of samples in CXR images (pneumonia) data set.

Focus of infection	Samples
Pneumonia	4273
Normal	1583

Data set source: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

In order to further verify the versatility of this model for the identification of pneumonia in CXR based on the transformer backbone network, a comparative experiment on the CXR Images (Pneumonia) dataset [21] was carried out and compared its result with the experimental results of other models on the same dataset.

Finally, in the experiment, the decision factors of the identification in the chest X-ray image from the Swin transformer were extracted, and with the Grad-CAM [23] they were superimposed on the original image to perform the discriminative output of the lesion area.

3. Experimental Result

The accuracy of the experiment results with the model in this paper on the Chest X-ray data set reached 87.3%. From the comparison in Table 3, it can be seen that the model based on the Swin Transformer backbone network and optimized for CXR is obviously better than other models based on traditional convolutional neural network.

In the experiment to verify the effectiveness of preprocessing of image enhancement according to the features of CXR, this paper collects the accuracy data of the model based on the Swin Transformer backbone network during the training process. As shown in Figure 3, in order to show the details more clearly, the figure draws a line graph of the accuracy from batches 32 to 128 on the first epoch without image enhancement and with image enhancement, in which the blue dotted line is no enhancement, and the orange

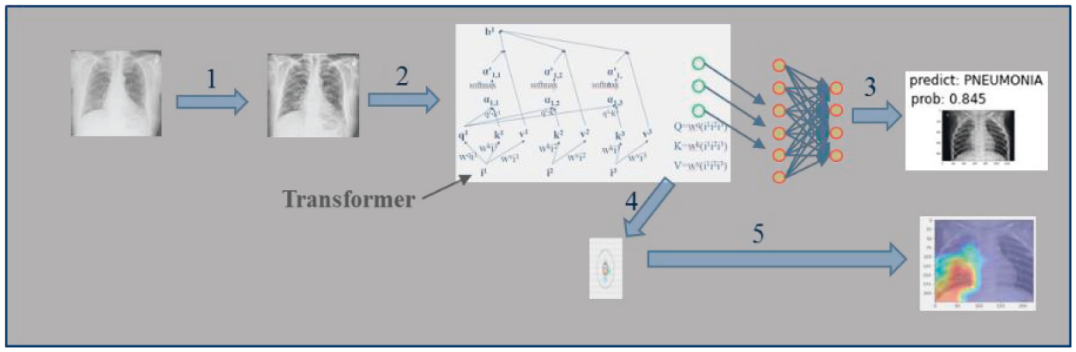
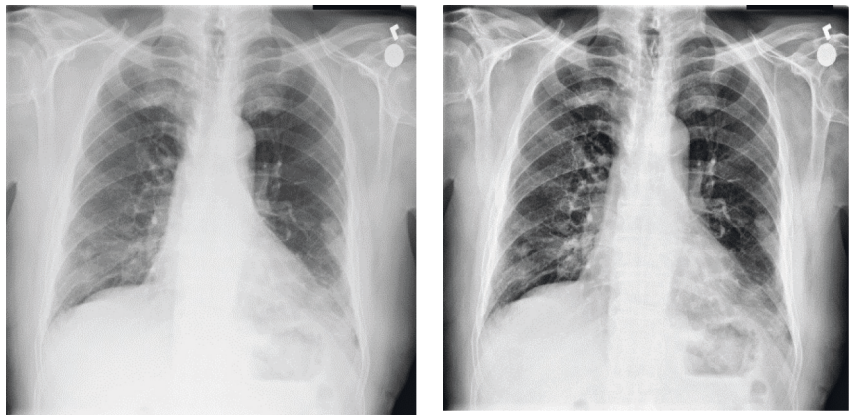


Figure 1: Schematic of a transformer network for CXR.



(a) (b)

Figure 2: Image enhancement (a) is before enhancement and (b) is after enhancement.

Table 3: Comparison of the models based on different backbone network on data set 1.

	Backbone network	Validate-accuracy (%)
Wang, et al. [9]	AlexNet	54.9
Wang et al. [9]	GoogLeNet	59.9
Wang et al. [9]	VGGNet-16	51.0
Wang et al. [9]	ResNet-50	63.3
Yao et al. [10]	DenseNet	71.3
Rajpurkar et al. [11]	DenseNet-121	76.3
This paper	SwinTransformer	87.3

dashed line is enhancement. As can be seen from Figure 3, the accuracy of image enhancement according to the features of CXR will be higher than that without image enhancement under the same circumstances.

In the comparative experiment on the CXR Images (Pneumonia) data set, the model based on the Swin Transformer backbone network and optimized for CXR in this paper achieved the best accuracy of 97.2% after only five epochs of training, which is much higher than the accuracy rate of 92.8% from the model based on the convolutional neural network proposed by Kermany's team [13]. It is also higher than the competition results in the Kaggle on the CXR Images (Pneumonia) data set (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/discussion/). The comparison data of accuracy from different models are shown in Table 4.

In Figures 4 and 5, the cross-entropy loss and accuracy during the training process using the model in this paper are shown. The top figure shows the change of the cross-entropy loss on epochs (the blue dotted line is from the data of the training set, and the orange dashed line is from the data of the validation set), and the bottom figure shows the change of the accuracy on epochs (the blue dotted line is from data of the training set, and the orange dashed line is from the data of the validation set).

What is the reason for the higher accuracy on the validation set than on the corresponding training set (Figure 5)? Because in order to enhance the generalization ability of the model, the data of the training set are randomly scaled, cropped, and flipped before entering the transformer network to extract features, while the data of the validation set has not undergone this transformation.

Before the transformation of the Softmax function and entering the fully connected classification network, the decision factors of the identification in chest X-ray image from the transformer are extracted. In our experiments, the decision factors are from the norm layer following the transformer backbone network, which can be obtained by back-propagating the result value of the latter classification network. The reverse derivation according to the Grad-CAM

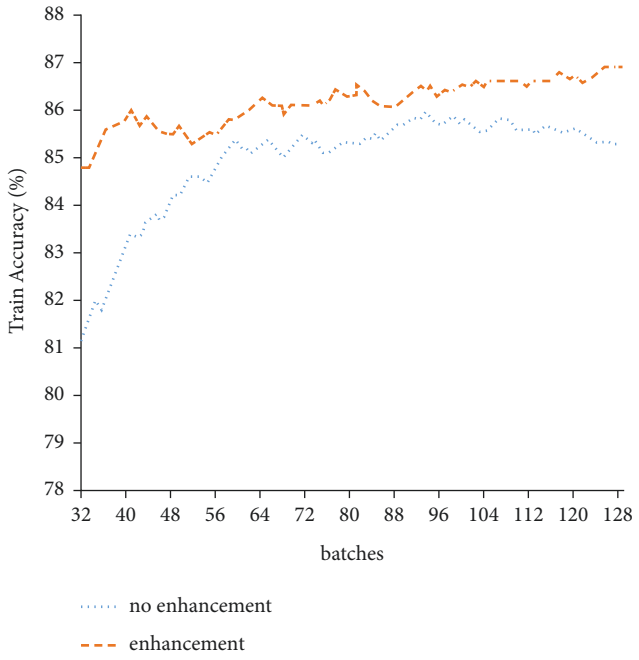


Figure 3: Comparison accuracy between no enhancement and enhancement on 1st epoch in the training process.

Table 4: Comparison of the models based on different backbone network on data set 2.

	Backbone network	Validate-accuracy (%)
Sharma et al. [13]	—	92.8
Grzegorz on Kaggle*	ResNet	94
This paper	SwinTransformer	97.2

*https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/discussion/313883.

algorithm is superimposed with the original image to form a heat map as shown in Figure 6. The area with high color temperature is the area that plays an important role in the formation of network discrimination, so the corresponding lesion area can be obtained.

4. Discussion and Analysis

After several years of research, the research of the application models based on the traditional convolutional neural backbone networks such as AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet, and EfficientNet has gradually entered the bottleneck period of network performance improvement, and the improvement effect in the application models research of CXR also gradually becomes less obvious. At this time, a new backbone network is urgently needed to solve this problem.

Vaswani et al. from the Google team proposed the transformer backbone network in 2017 [24]. Compared with the traditional Recurrent Neural Network [25], Transformer has many advantages such as infinite memory length in theory and parallel operation. The theory of the self-attention algorithm is the basis of the transformer (1).

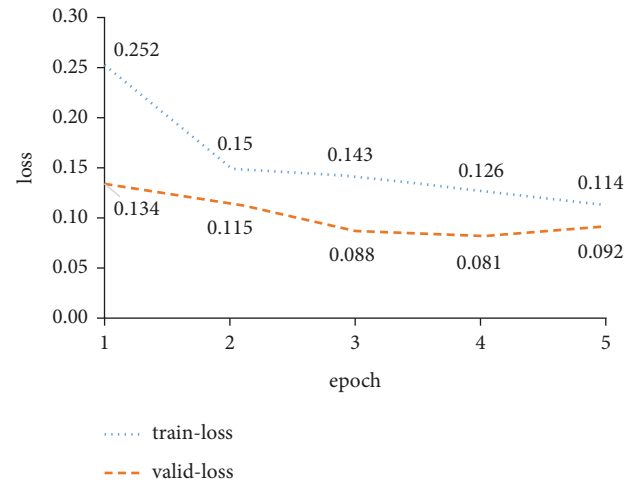


Figure 4: Cross-entropy loss in the training process and validation process.

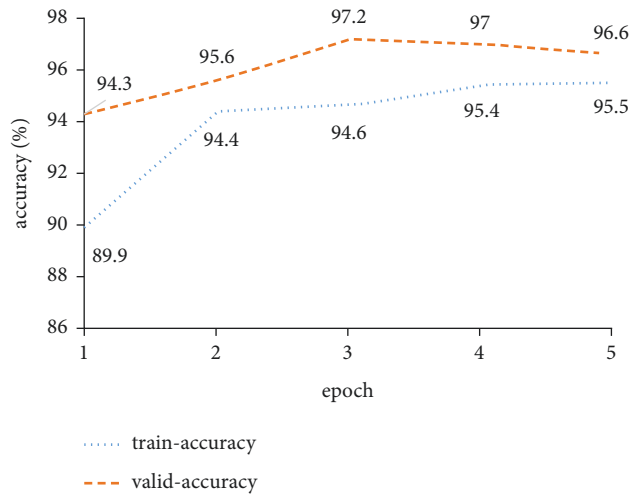


Figure 5: Accuracy in the training process and validation process.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

In transformer multihead, self-attention extended from the self-attention algorithm is used, and it is split by linear mapping according to the number of headers and is usually divided equally (2).

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O,\\ \text{where head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right),\\ \text{Attention}(Q, K, V) &= \text{Softmax}\left(\frac{QK^T}{\sqrt{.d_k}}\right)V.\end{aligned} \tag{2}$$

The transformer was originally used for natural language processing. At the 2020 International Conference on Computer Vision and Pattern Recognition (CVPR), the

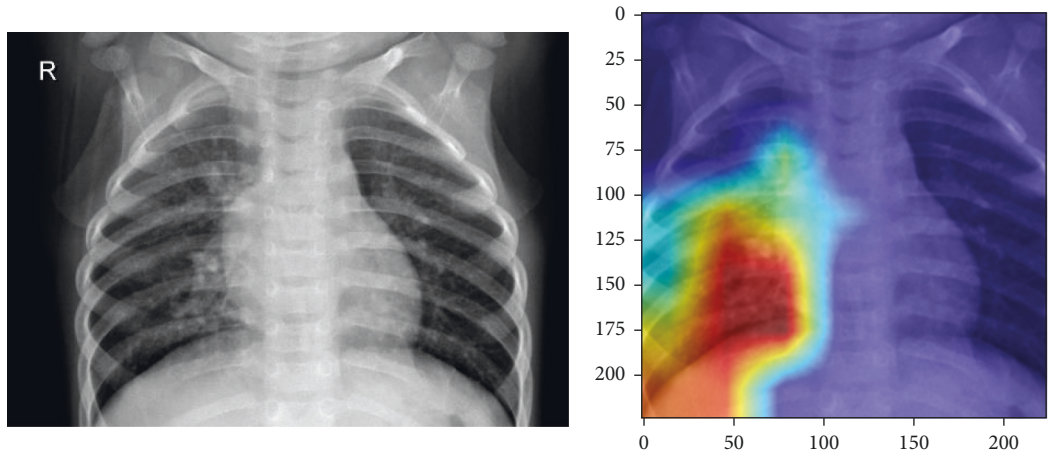


Figure 6: Heat map for lesion area.

Google team proposed a scheme to apply transformers to the field of computer vision and achieved good results [26]. The core of the transformer comes from the self-attention algorithm, and the self-attention algorithm and the convolution algorithm are very closely related, and the latter can be considered as a subset of the former [27]. As the scale of the data set increases, the performance of the transformer backbone network will exceed the traditional convolutional neural backbone networks, and large-scale training data can encourage the transformer to learn the more translation equivariance and locality than possessed by the convolutional neural networks. In 2021, the Swin Transformer proposed by Microsoft Research Asia has become a bright spot. It overcomes the bottleneck of the traditional convolutional neural backbone network to a certain extent and further improves the accuracy [22]. The experimental comparison results [22] of its effect on the ImageNet-1K data set are as follows shown in Table 5. The difference from convolutional neural networks such as ResNet is that the Swin Transformer no longer uses traditional convolution kernels in feature extraction, and the core at each level is window multihead self-attention and shifted window multihead self-attention. Window multihead self-attention is the multihead self-attention matrix operation performed inside the window. The advantage is to reduce the amount of computation, but the disadvantage is that information interaction between windows is not possible. The shifted window multihead self-attention is designed to overcome this shortcoming, and it can realize the information interaction between different windows by shifting the window position.

Since the Swin transformer backbone network has not been proposed for a long time, its application in various fields has not been sufficiently studied, and few studies have been conducted to optimize the model for the application of CXR images. Regarding the application model in the CXR field, the model proposed in this paper is no longer based on the traditional convolutional neural backbone network, but the Swin transformer backbone network is introduced to build the model. The experimental results on the two CXR data sets prove that the accuracy of the model based on the transformer backbone network is higher than that of the model based on the traditional convolutional neural backbone network model on the same data set, and the effect is improved obviously, which can overcome the existing bottleneck of improving the accuracy of model based on the convolutional neural backbone network.

Table 5: Comparison of different backbone network.

	Throughput (images/s)	Accuracy (%)
EfficeintNet-B3	732.1	81.6
EfficeintNet-B4	349.4	82.9
Swin-T	755.2	81.3
Swin-S	436.9	83.0
Swin-B	278.1	83.5

It is necessary to highlight image details and suppress noise because CXR is usually characterized by low brightness, low contrast, and large noise. In the histogram of CXR, the area with the highest pixel distribution is usually the background, which is a nonconcern area, so this part can be peaked. If the values in the histogram are evenly distributed, it means that the distribution on each gray level is balanced, and the contrast is the best at this time, that is, the image is generally clear. Therefore, a certain degree of equalization processing on each gray level in the histogram of CXR is helpful to the subsequent processing. It can be seen from the data chart of the experimental results (Figure 3) that this processing method is effective for the model based on the transformer backbone network.

The concept of class activation mapping (CAM) originated from the interpretability research of deep neural networks [28] and was later introduced into application research by some scholars. On this basis, Selvaraju et al. proposed gradient weighted activation mapping (Grad-CAM) [29].

In the experiment of CXR in this paper, gradient weighted activation mapping is combined into the transformer, so that by extracting the decision factors of the identification in CXR from the transformer, the heat map through reverse derivation is superimposed on the original

image, so the corresponding approximate lesions area can be found.

When the new type of pneumonia caused by the 2019 novel coronavirus appeared, the study of the corresponding model was carried out. Narin et al. used ResNet50, ResNet101, ResNet152, Inception V3, and Inception-ResNet-V2 five models for identification of the new type of pneumonia and compared them. Their experimental results also show that ResNet50 achieves better results [30]. In similar cases, it is estimated that the model with the Transformer backbone network will be better.

5. Conclusion

Pneumonia is a disease with a high mortality rate. Chest X-ray imaging is widely used in the routine examination of pneumonia. CXR as an important adjunct to the diagnosis of pneumonia can diagnose pneumonia quickly and accurately. Machine learning methods based on deep learning have been effective in chest X-ray imaging. In this paper, the Swin Transformer is applied to the application model of CXR image recognition and analysis, and the model is optimized accordingly according to the characteristics of CXR. The experimental results show that the model outperforms the model based on the traditional convolutional neural backbone network.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant no. 2021A1515310003).

References

[1] S. L. Murphy, J. Xu, K. D. Kochanek, S. C. Curtin, and E. Arias, "Deaths: final data for 2015. National vital statistics reports: from the centers for disease control and prevention, national center for health statistics," *National Vital Statistics System*, vol. 66, pp. 1–75, 2017.

[2] K. Y. Yoon, S. H. Joo, K. M. Joon, and L. M. Jung, "Comparison of radiation dose from X-ray, CT, and PET/CT in paediatric patients with neuroblastoma using a dose monitoring program," *Diagnostic and Interventional Radiology*, vol. 22, no. 4, pp. 390–394, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, "Going Deeper with Convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Boston, MA, June 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] G. Huang, Z. Liu, V. Laurens, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society. IEEE Computer Society, Honolulu, HI, USA, July 2017.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019, https://arxiv.org/pdf/1905.11946.pdf.11946.

[9] 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," in *Proceedings of the IEEE. Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3462–3471, Honolulu, HI, USA, December 2017.

[10] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to Diagnose from Scratch by Exploiting Dependencies Among Labels," 2017, https://arxiv.org/abs/1710.10501.

[11] P. Rajpurkar, J. Irvin, K. Zhu et al., "Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning," 2016, https://arxiv.org/abs/1711.05225.

[12] S. A. Khoiriyah, A. Basofi, and A. Fariza, "Convolutional neural network for automatic pneumonia detection in chest radiography," in *Proceedings of the 2020 International Electronics Symposium (IES)*, Surabaya, Indonesia, September 2020.

[13] H. Sharma, J. S. Jain, P. Bansal, and S. Gupta, "Feature extraction and classification of chest X-ray images using CNN to detect pneumonia," in *Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) IEEE.*, Noida, India, January 2020.

[14] I. Masad, A. Alqudah, A. M. Alqudah, and S. Almashaqbeh, "A hybrid deep learning approach towards building an intelligent system for pneumonia detection in chest X-ray images," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, Article ID 5530, 2021.

[15] Z. Parveen, S. Adinarayana, R. Aamani, S. Santoshi, and S. Tulasi, "Efficient pneumonia detection in chest xray images using convolution neural network," *International Journal of All Research Education and Scientific Methods*, vol. 9, no. 7, pp. 2900–2905, 2021.

[16] R. K. Gupta, Y. Sahu, N. Kunhare, A. Gupta, and D. Prakash, "Deep learning based mathematical model for feature extraction to detect corona virus disease using chest X-ray images," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 29, no. 06, pp. 921–947, 2021.

[17] R. Alsharif, Y. Alissa, A. M. Alqudah, I. A. Qasmieh, A. M. Wan, and H. Alquran, "PneumoniaNet: Automated Detection and Classification of Pediatric Pneumonia Using Chest X-ray Images and CNN Approach," *Electronics*, vol. 10, no. 23, pp. 1–13, 2021.

[18] W. Xue, "Classification of pulmonary lesions based on CNN and chest X-ray images," *Journal of Physics: Conference Series*, vol. 1952, no. 2, Article ID 022025, 2021.

[19] M. J. Alam, S. N. Ali, and M. Z. Hasan, "A Robust CNN Framework with Dual Feedback Feature Accumulation for Detecting Pneumonia Opacity from Chest X-ray Images," in *Proceedings of the 2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, Dhaka, Bangladesh, December 2020.

[20] C. Han, T. Okamoto, K Takeuchi et al., "Tips and Tricks to Improve CNN-Based Chest X-ray Diagnosis: A Survey," 2021, https://arxiv.org/abs/2106.00997.

[21] D. S. Kermany, M. Goldbaum, W. Cai et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[22] Z. Liu, Y. Lin, Y. Cao et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," 2021, https://arxiv.org/abs/2103.14030, Article ID 14030.

[23] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: why did you say that?," 2016, https://arxiv.org/abs/1611.07450.

[24] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention Is All You Need," 2017.

[25] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *Computer Science*, 2015.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby, "An Image Is worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, https://arxiv.org/abs/2010.11929.

[27] J. B. Cordonnier, A. Loukas, and M. Jaggi, "On the Relationship between Self-Attention and Convolutional Layers," 2020, https://arxiv.org/abs/1911.03584.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[30] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (Covid-19) Using X-ray Images and Deep Convolutional Neural Networks," *Pattern Anal Appl*, vol. 24, no. 3, 2020.