Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 1927860, 13 pages
https://doi.org/10.1155/2021/1927860

Research Article

Research on Intelligent Recognition Algorithm of Pneumonia Based on Deep Convolution and Attention Neural Network

Qiongqin Jiang,[1] Wenguang Song,[1] Gaoming Yu,[1] Ming Zhao,[1] Bowen Li,[1] Haoyuan Li,[1] and Qian Yu[2]

[1] *Cooperative Innovation Center of Unconventional Oil and Gas, Yangtze University, Ministry of Education, Hubei Province, Wuhan 43400, China*
[2] *Faculty of Engineering and Applied Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada*

Correspondence should be addressed to Wenguang Song; wenguang_song@yangtzeu.edu.cn and Gaoming Yu; ygm1210@vip.sina.com

Received 23 June 2021; Revised 14 August 2021; Accepted 24 August 2021; Published 13 September 2021

Academic Editor: Nianyin Zeng

Pneumonia is a common infection that inflames the air sacs in the lungs, causing symptoms such as difficulty breathing and fever. Although pneumonia is not difficult to treat, prompt diagnosis is crucial. Without proper treatment, pneumonia can be fatal, especially in children and the elderly. Chest x-rays are an affordable way to diagnose pneumonia. Investigating an algorithmic model that can reliably and intelligently classify pneumonia based on chest X-ray images could greatly reduce the burden on physicians. The advantages and disadvantages of each of the four convolutional neural networks VGG16, ResNet50, DenseNet201, and DWA algorithm models are analyzed and given by comparing and investigating each model. The VGG16, ResNet50, and DenseNet201 network models are compared with the DWA model. When training the depthwise separable convolution with attention neural network (DWA), the training accuracy reaches 97.5%. The validation accuracy was 79% due to the model's tendency to overfit, and the test dataset had 1175 X-ray images with a test accuracy of 96.1%. The experimental results illustrate the effectiveness of the attention mechanism and the reliability of the deeply separable convolutional neural network algorithm. The successful application of the deep learning algorithm proposed in this paper on pneumonia recognition will provide an objective, accurate, and fast solution for medical practitioners and can provide a fast and accurate pneumonia diagnosis system for doctors.

1. Introduction

The intelligence of deep learning algorithms is due to the fact that they use objective criteria and take advantage of the fast computational speed of computers to make judgments. Over the years, deep learning algorithms have achieved world-renowned research results in various fields. The amount of data in the biomedical field is very large and complex, the data types are diverse, and the meaning of the data is often not easy to understand. Using deep learning methods to analyze and process biomedical data can often be very effective. In recent years, a consensus has emerged between the computer and biomedical fields to adopt an AI with medical approach to such tasks. In this area, research has been conducted on how to use AI algorithms to deal with various biomedical problems. For example, AI techniques can be applied to medical image classification tasks to intelligently identify medical images. Computers search databases based on image features, combine image features with clinical report judgments, and train models on large sets of collected medical data or complete sets of genetic sequences to obtain and identify individual features of each person. The recognition of medical X-ray images using deep learning algorithms is an innovative research in the field of pneumonia diagnosis. The lung X-ray image dataset used in this paper includes the entire left lung and right lung, which can be used to evaluate the overall and local lesions of the lung and thus make effective judgments. Normal or abnormal pneumonia images are determined by comparing the measured pneumonia X-ray images with normal X-ray images to accurately diagnose pneumonia.

2. Pneumonia Image Recognition Model Design

In image recognition, the structure of the deep convolutional neural network is used in this paper, of which there are four types of convolutional neural networks: VGG16, ResNet50, DenseNet201, and DWA. It intuitively shows the recognition accuracy of excellent convolutional neural networks in the ImageNet industry competition in recent years [1], and the comparison of the number of network model parameters is given in Figure 1.

2.1. VGG Model. The VGG model is characterized by the use of multiple smaller convolutional filters instead of one larger convolutional filter, and by this operation the network depth is deepened, thus improving the learning ability of the neural network [2]. The advantage of VGG16 is that the structure is very simple, the size of the convolutional kernel used for the whole network is (3×3), and the maximum pooling size is (2×2). VGG16 has 3 fully connected layers, so its disadvantage is that it requires more computational resources and uses more parameters, most of which come from the first fully connected layer [3]. Thus VGG16 has more memory usage (140M). The architecture of the VGG16 network is shown in Figure 2.

As shown in Figure 3, the VGG model uses three 3×3 convolutional filters instead of 7×7 convolutional filters, which not only reduces the number of network model parameters but also makes the network more suitable for handling nonlinear complex tasks.

2.2. ResNet Model. Residual neural network (ResNet) was proposed by Similä et al., [4], Zhang Xiangyu, Ren Shaoqing, Sun Jian, and others of Microsoft Research [5]. ResNet won the championship in the 2015 ILSVRC (ImageNet Large Scale Visual Recognition Challenge) [6].

He Kaiming proposed a residual learning method to solve the degradation problem [7]. For a stacked layer structure (consisting of several layers), when the input is x, the learned features are denoised as $H(x)$. Now we want it to know that the residual $F(x)$ is equal to $H(x) - x$, so the original learned features are actually $F(x) + x$. When the residual value is zero, the stacked layers only perform identity mapping, which at least does not degrade the network performance. The fact that the residuals are not zero also allows the stack layer to learn new features on top of the input features, resulting in better performance [8]. Thus, learning residuals is easier than learning the original features directly.

The main contribution of the residual neural network is the invention of the "shortcut connection" for the degeneracy phenomenon, which greatly eliminates the problem of difficulty in training neural networks with too much depth. The "depth" of neural networks exceeded 100 layers for the first time, and the largest neural networks even exceeded 1000 layers. The phenomenon of degradation refers to the fact that as the layers of the network deepen, the accuracy of the model first increases, reaches a maximum (accuracy saturation), and then unpredictably decreases significantly as the depth of the network continues to increase [9]. The structure of residual learning is shown in Figure 4.

$$\begin{aligned} y_l &= h(x_l) + F(x_l, W_l), \\ x_{l+1} &= f(y_l). \end{aligned} \tag{1}$$

In equation (1), x_l and x_{l+1} represent the input and output of the l residual unit, respectively, and each residual unit usually contains a multilayer structure. $h(x_l) = x_l$ represents the constant mapping, F is the residual function, y represents the learned residuals, and f is the ReLU activation function. Based on the above equations, we derive the learning features from the shallow l to the deep L layers as follows:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i). \tag{2}$$

Using the chain rule, the gradient of the back-propagation process can be found.

$$\frac{\partial \text{loss}}{\partial x_l} = \frac{\partial \text{loss}}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial \text{loss}}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i)\right), \tag{3}$$

where $\partial \text{loss}/\partial x_L$ denotes the gradient when the value of the loss function reaches L and 1 denotes the lossless propagation gradient. The other residual gradients need to pass through the layers of the entitled parameters rather than directly through the gradient. The gradient of the residuals is not always -1, and the presence of 1 does not make the gradient disappear, even if the gradient is small. Therefore, learning the residuals is easier than learning the original features directly.

The ResNet network is a network improvement based on VGG19 [10]. The residuals are obtained by a short-circuiting mechanism, as shown in Figure 5. This change is mainly reflected in the fact that ResNet directly uses convolution with stride = 2 for subsampling and uses a global average pooling layer instead of a fully connected layer [11]. A key design rule of ResNet is to double the number of feature maps (feature maps) when their size is reduced to half the size in order to maintain the complexity of the network layers. As can be seen in Figure 5, the short-circuiting mechanism is added between every two layers to form residual learning in RESNET compared to a normal neural network, where the dashed line indicates the change in the number of feature maps. 34 layers of ResNet are shown in Figure 5, and deeper networks can also be built. When the network is deeper, the residuals are learned between the three layers. The three layers of convolution kernel are 1×1, 3×3, and 1×1. It is 1/4 of the output feature map. The important design principle of ResNet50 is to keep the complexity of the network layers constant while halving the feature map size and doubling the number of feature maps [12].

ResNet uses two residual units, as shown in Figure 6. The image on the left is the shallow network and the image on the right is the deep network. When the input and output

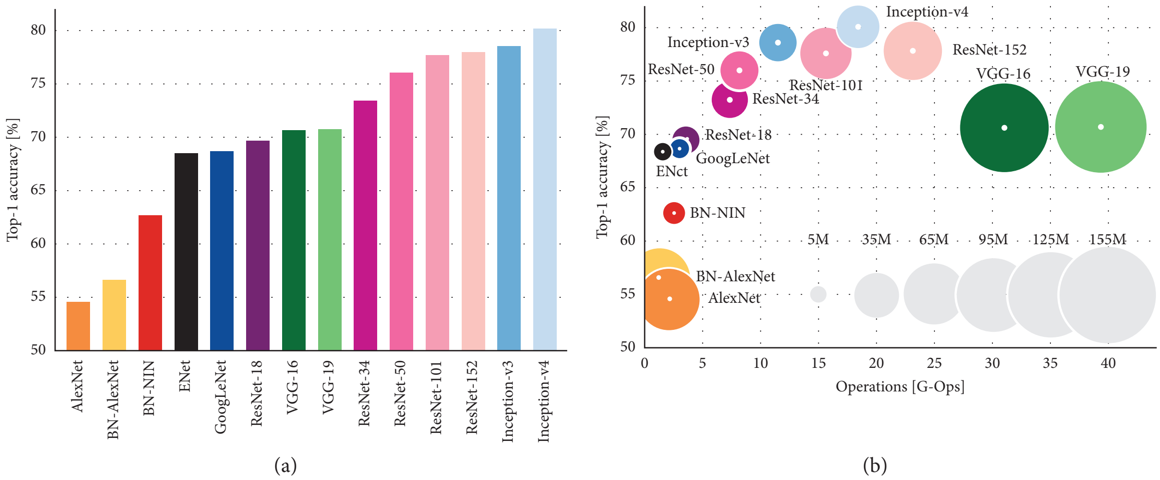


Figure 1: The comparison of various excellent depth convolutional models.

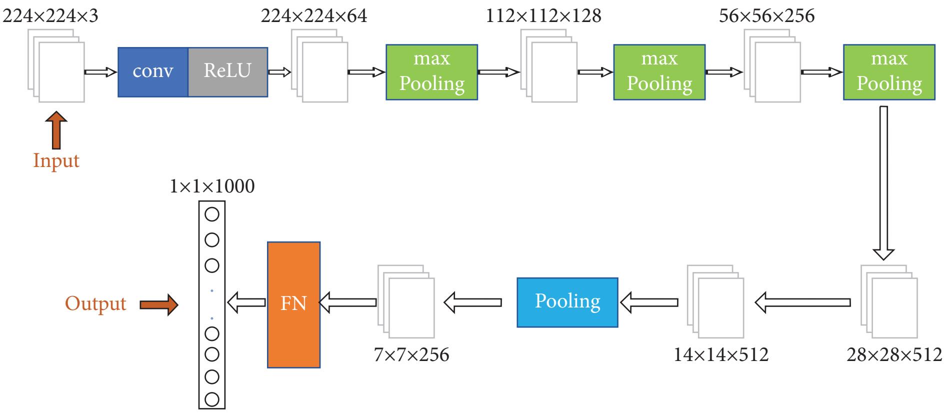


Figure 2: VGG16 network architecture diagram.

dimensions are the same, a short-circuit connection can directly add the input to the output [13]. However, when the dimensions are different (equivalent to doubling the dimensions), they cannot be added directly. There are two methods. The first method, zero-padding, is used to add dimensions. In this case, a downsampling (downsamp) operation is usually performed first, followed by a step (stride) = 2 pooling operation to keep the number of parameters at the original size [14]. In the second approach, a new mapping method (projection shortcut) is used. The operation is generally performed using 1 × 1 convolution, but this increases the number of parameters and computation.

A comparison of the effect of 18-layer and 34-layer networks is shown in Figure 7. It can be seen that the normal network shows the degradation of training effect, while ResNet solves the training effect degradation problem very well.

2.3. DenseNet Model. The DenseNet model proposes a dense connectivity mechanism, that is, a mechanism for interconnecting the layers. Specifically, each layer will accept all previous layers as its additional input. This model allows convolutional networks to be trained more deeply, accurately, and efficiently. ResNet is a short-circuit connection between each layer and the previous layer (usually 2 to 3 layers) by component-level summation. In DenseNet, each layer is connected to all previous layers in the channel dimension (each layer has the same size feature map), which together serve as the input to the next layer. For an L layer network, DenseNet contains a total of $L(L+1)/2$ connections; compared to ResNet, DenseNet is a dense connection. DenseNet connects all layers directly while ensuring maximum information transfer between layers in the network, enhancing the transfer of features, using them more efficiently, and reducing the number of parameters to some extent. Due to the excellent structural design of the DenseNet model, the feature reuse is enhanced, and the gradient dispersion problem is better solved, while the number of network parameters is greatly reduced. The output of the traditional network at the l layer is as follows:

$$x_l = H_l(x_{l-1}). \tag{4}$$

The ResNet network adds the identity function from the input of the previous layer as follows:

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input (224 * 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

FIGURE 3: VGG model structure diagram.

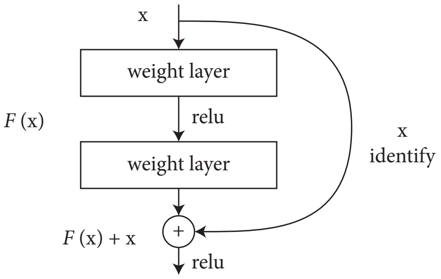


FIGURE 4: Residual learning unit.

$$x_l = H_l(x_{l-1}) + x_{l-1}. \quad (5)$$

The DenseNet network connects all previous layers as input.

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \quad (6)$$

where $H_l(\cdot)$ represents the nonlinear transformation function, which can consist of a series of BN (batch normalization), ReLU, pooling, and convolution operations. In fact, there may be more than one convolutional layer between layers l and $l-1$.

CNN networks generally reduce the size of feature maps by pooling or convolution with step >1, while the densely connected model of DenseNet requires the feature maps to be of the same size. To solve this problem, the DenseBlock + transition structure is used in the DenseNet network. DenseBlock is a module consisting of multiple layers, each with the same size feature map, and the layers are densely connected to each other. The transition module connects two adjacent dense blocks and reduces the size of the feature map by pooling. The network structure of DenseNet is shown in Figure 8. It consists of four dense blocks which are connected together by transitions [15].

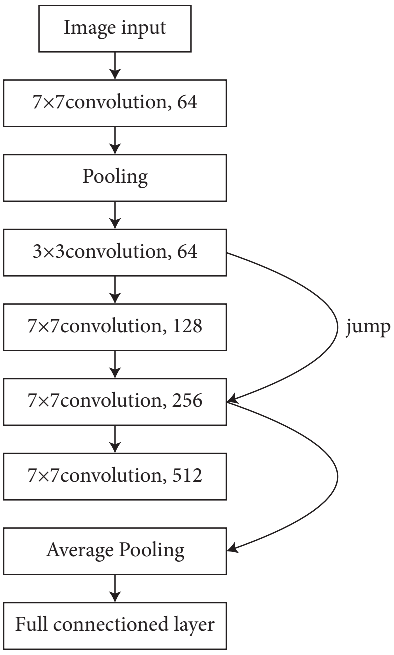


FIGURE 5: ResNet network structure.

As shown in Figure 8, the nonlinear combinatorial function $H(\cdot)$ in DenseBlock consists of the structure BN + ReLU + 3 × 3 Conv. It is worth mentioning that, unlike ResNet, all layers in DenseBlock are convolved with k feature map outputs, that is, the number of channels of the resulting feature map is k; that is, k convolution kernels are used. In DenseNet, it is called growth rate, which is a hyperparameter k. In most cases, a smaller k (e.g., a value of 12) is used to obtain better performance. Assuming that the number of channels in the feature map of the input layer is k_0, the number of channels in the input of layer l is $k_0 + k(l-1)$. As the number of layers increases, the input of the DenseBlock becomes larger due to feature reuse, even if k is set small. Only k features are unique for each layer.

Due to the large number of input parameters at the bottom layer, DenseBlock can adopt the bottleneck layer to reduce the computation, which adds 1 × 1 Conv, BN + ReLU + 1 × 1 Conv + BN + ReLU + 3 × 3 Conv to the original structure, called DenseNet-B structure. The 1 × 1 Conv can obtain4k feature map, which can reduce the number of features and improve the computational efficiency.

The role of the transition layer is to reduce the size of the feature map, which consists of 2 × 2 AVGpooling and Bn + ReLU + 1 × 1 Conv + 2 × 2 AVGpooling structure of

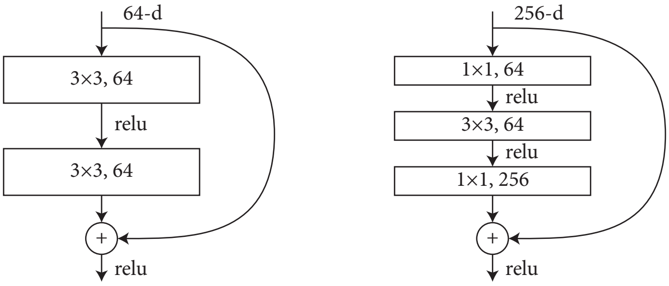


Figure 6: Different residual unit.

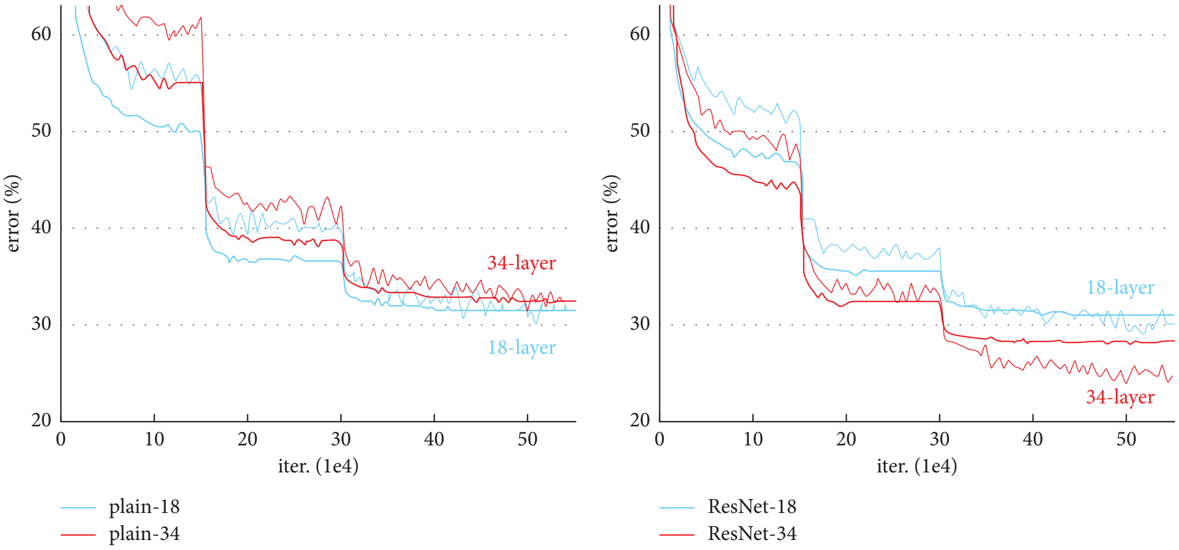


Figure 7: Comparison of network effects between ResNet-18 and ResNet-34.

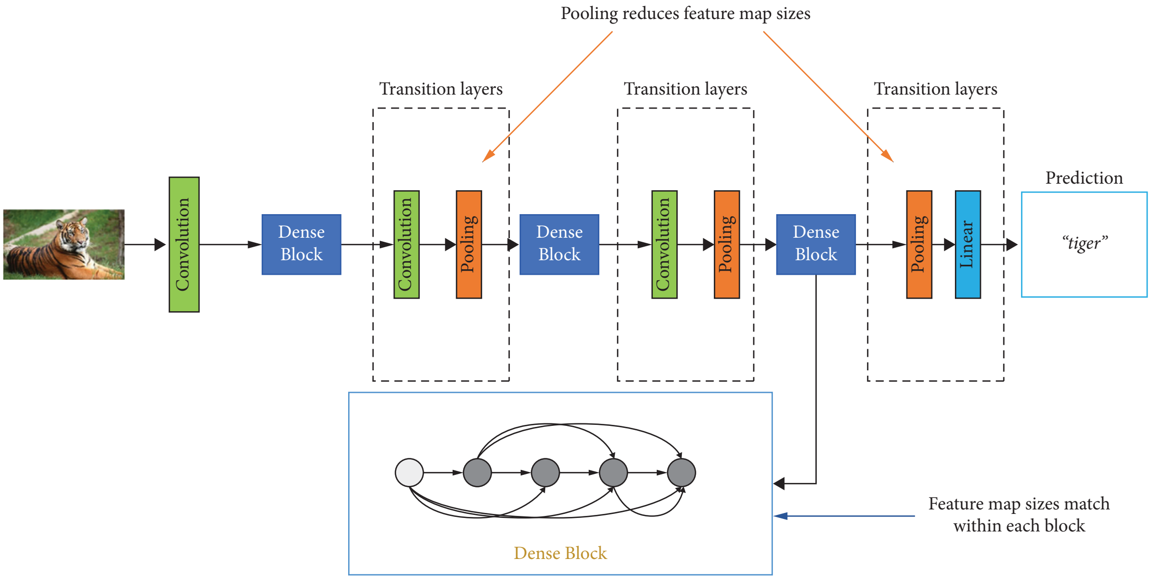


Figure 8: DenseNet model structure.

1×1 convolution; such a structure can compress the network model. Assume that the number of channels in the feature map obtained by connecting the DenseBlock on the transition layer is m. The transition layer can produce $m * \theta$ features (through the convolution layer), where $\theta \in (0, 1)$ is the compression rate. The number of features passing through the transition layer does not change; that is, there is no compression. When the compression coefficient is less than 1, this structure is called densenet-c. The combined structure of DenseBlock structure using a bottleneck layer and a transition with a compression coefficient less than 1 is called DenseT-BC.

2.4. DWA Model. DWA (depthwise separable convolution with attention neural network) is a depthwise separable convolutional network based on attention mechanism [16]. Depthwise Separable Convolution (DSC) first appeared in a PhD thesis entitled "Rigid-motion scattering for image classification." The detailed structure of DSC is shown in Figure 9. The depthwise separable convolution consists of two parts: Depthwise Convolution and Pointwise Convolution. The computation of Depthwise Convolution is very simple. It uses one convolution kernel for each channel of the input feature map and then stitches the output of all the convolution kernels to get its final output, as shown in the Depthwise Convolution part in Figure 9. The depth convolution section is shown.

Since only one convolutional kernel is used for each channel in a deeply separable convolutional network, the number of output channels of the convolutional operation is the number of convolutional kernels. So if the number of channels in the input feature map is N, the number of channels in the input feature map is 1, and the number of channels in the input feature map is 1. Then the N feature maps are stitched together in order to obtain an output feature map with N channels [17].

The point-by-point convolution is actually a 1×1 convolution, which plays two roles in the depth-separable convolution. The first role is to allow the depth-separable convolution to freely change the number of output channels; the second role is to perform channel fusion on the feature map output from the depth convolution. The first role is easier to understand, because the depth convolution alone cannot change the number of output channels, and thus it is more intuitive and simple to use 1×1 convolution to change the number of output channels. To understand the second role of point-by-point convolution, consider what happens when only deep convolution is used to stack the network. Suppose that the input is IN and its i-th channel is denoted as IN_i; the output of the first layer of deep convolution is denoted as DC1 and its i-th channel is denoted as DC1_i; the output of the second layer of deep convolution is denoted as DC2 and its i-th channel is denoted as DC2_i [18].

From the working mechanism of deep convolution, it is clear that DC1_i is only related to IN_i, DC2_i is only related to DC1_i, and in turn DC2_i is only related to IN_i. Simply put, there is no computation linking the different channels of input and output together [19]. The $1 * 1$ convolution itself has the ability of channel fusion, so the point-by-point convolution followed by deep convolution can effectively solve the above problems.

Suppose that the size of the input feature map is $D_k \times D_k \times M$, the size of the convolution kernel is $D_F \times D_F \times M$, and its number is N. Assuming that one convolution operation is performed for each point in the corresponding feature map space location, it is known that a single convolution requires a total of $D_k \times D_k \times D_F \times D_F \times M$ computations. This is because the feature map space dimension contains a total of $D_k \times D_k$ points, and the computation of convolution operation for each point is the same as the size of the convolution kernel, which is $D_k \times D_k \times M$, so the total computation for a single convolution is as follows:

$$D_k \times D_k \times D_F \times D_F \times M. \tag{7}$$

Then, for N convolution, the total computation is as follows:

$$D_k \times D_k \times D_F \times D_F \times M \times N. \tag{8}$$

Using a similar analysis for the depth-separable convolution shows that the total amount of depth convolution computation is as follows:

$$D_k \times D_k \times D_F \times D_F \times M. \tag{9}$$

The total amount of point-by-point convolution is as follows:

$$M \times N \times D_K \times D_K. \tag{10}$$

Therefore, the total calculated amount of deep separable convolution is as follows:

$$D_k \times D_k \times D_F \times D_F \times M + M \times N \times D_K \times D_K. \tag{11}$$

Then, relative to the ordinary convolution, the ratio of the calculation amount of the deep separable convolution to the ordinary convolution is as follows:

$$\frac{1}{N} + \frac{1}{D_F^2}. \tag{12}$$

As can be seen above, the computational efficiency of depth-separable convolution is much better than that of ordinary convolution.

The standard convolutional layer to acquire features is a standard method in many computer vision tasks. In an encoding-decoding network, the input image is convolved, activated, and pooled to obtain a feature vector, which is then recovered to an output image of the same size as the input image [20]. This method consists of specially designed convolutional blocks. Due to its simplicity and accuracy, this architecture is widely used. However, encoding-decoding networks have their limitations. As an example, a 3×3 convolutional filter computes 9 pixels, that is, computes the value of the target pixel with reference to itself and the surrounding 8 pixels only. It can only use the local information to calculate the target pixel and ignore the global information. Although this drawback can be mitigated by some simple methods, that is, using larger

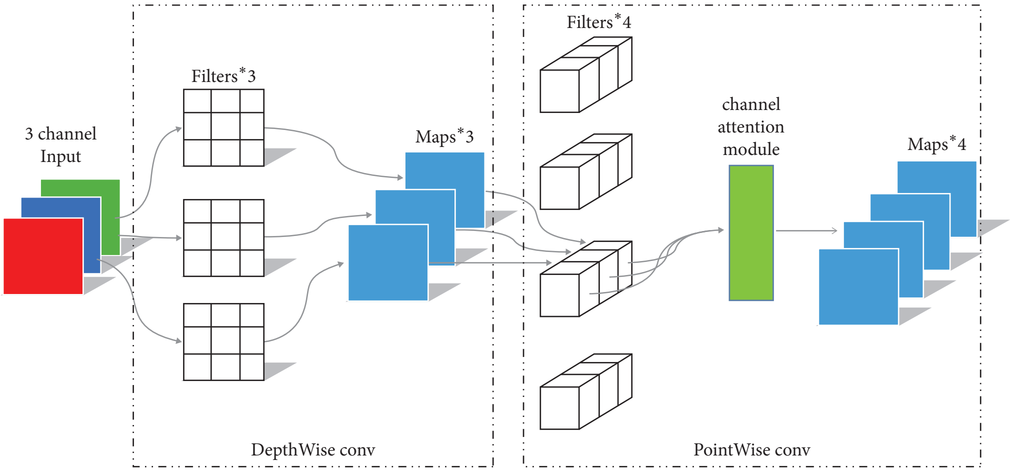


Figure 9: DWA convolution structure.

convolution kernels or employing more convolution layers, this would result in greater arithmetic power consumption and no significant improvement in the drawback. The principle of the attention mechanism is to treat each pixel as a random variable and compute the covariance between all two-two pixels. The target pixels involved are the weighted sum of all pixel values; that is, the global reference information is obtained by weighting. The value of each predicted pixel is enhanced or diminished according to the magnitude of the weights. Using similar features and ignoring dissimilar features during training and prediction can achieve the goal of obtaining global information while reducing the computational effort. This network uses the channel attention mechanism to obtain global information. The basic operation of the channel attention mechanism is described below.

The Squeeze operation turns each 2D feature channel into a channel with a global field of view to some extent, and the output dimension of the channel is equal to the number of input feature channels. This is useful in many tasks. The specific operation is to compress the batch size $*C*H*W$ into a batch size $*C*1*1$ using global average pooling [21].

The Excitation operation is used to generate weights for each feature channel through parameter W. The weights represent the correlation between the feature channels. Batch size $*C*1*1$ vector is input to a Bottleneck structure consisting of two fully connected layers, which serves to build the correlation between the network channels. The specific operation is to output the same number of weights as the input features. The feature dimension is first reduced to 1/16 of the input, then activated by ReLU [22], and then restored to the original dimension by the fully connected layers. Then the weights normalized between 0 and 1 are obtained by the Sigmoid function. The following is the mathematical representation of the channel attention mechanism [21].

First, note that the compression function of the module sums and averages all the eigenvalues of each channel, which is also the mathematical expression of the global average pooling. The equation of the compression function is as follows:

$$z_c = \mathbf{F}_{\text{sq}}(o_c) = \frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} o_c(m, n). \tag{13}$$

Next, the excitation function is used to activate each layer channel. The equation of the excitation function is shown as follows:

$$\mathbf{s} = \mathbf{F}_{\text{ex}}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})). \tag{14}$$

σ is the ReLU activation function and δ is the Sigmoid activation function. Two weights $(\mathbf{W}_1, \mathbf{W}_2)$ are learned by training to obtain a one-dimensional excitation weight to activate each layer of the channel [23].

Finally, the features of different channels are multiplied by different weights to focus on the key channel domain. The equation of the rescaled weight function is shown as follows:

$$\mathbf{x}_c = \mathbf{F}_{\text{scale}}(\mathbf{u}_c, s_c) = s_c \cdot \mathbf{u}_c. \tag{15}$$

The channel attention module of DWA performs a Squeeze operation and an Excitation operation on the feature maps to obtain the global features at the channel level. Then the weights of different channels are obtained by learning, and then the weights are multiplied with the original feature map to obtain the final features. The attention mechanism allows the model to focus more on the most informative channel features and suppress the unimportant channel features [24].

3. Analysis of Experimental Case

Three convolutional neural network models (ResNet50, VGG16, and DenseNet201) with excellent recognition results on the open-source dataset Chest X-Ray Images (Pneumonia) are experimented and the results are compared and analyzed with the innovative algorithm in this paper, the depthwise separable convolution with attention neural network (DWA), based on the attention mechanism. The results are compared and analyzed with the experimental results of depthwise separable convolution with attention neural network (DWA), an innovative algorithm in this paper. The effect of various optimization methods during network training on the recognition accuracy of the network models was tested. Finally, we combine the best-performing recognition models to construct an automatic medical image recognition system that can assist doctors in diagnosis. In this paper, all experiments are based on an NVIDIA GeForce RTX 2080Ti (12 GB) GPU to do data training and data testing. All models are learned using the Adam optimizer with an initial learning rate of 0.001. The environment configuration of the running system is Windows 10, TensorFlow2.0, and Python-3.8.4.

3.1. DWA Algorithm Framework. The automatic pneumonia recognition algorithm based on DWA is mainly divided into three steps, and the specific theoretical basis is introduced in the previous chapters. The specific implementation steps are as follows. First, lung X-ray images are acquired and preprocessed to build a deep learning classification network model framework. The original dataset is enhanced, and the enhanced dataset is fed into the network for training and prediction to obtain the classification results. The overall architecture is shown in Figure 10. Experiments were conducted according to the process shown in the overall framework to implement a DWA-based automatic pneumonia identification system.

3.2. Experimental Data and Image Enhancement Methods. There are fewer open-source lung X-ray image datasets due to the personal privacy involved in medical images. By comparing the open-source lung X-ray image datasets on the Internet, it is found that the public dataset of pneumonia on the Haggle platform has the best image quality. This dataset was selected from the chest radiographs of 1–5-year-old children in Guangzhou Women and Children Medical Center. The dataset was first quality-checked by experts on all its chest X-ray images, removing all low-quality (blurred) X-ray images or unrecognizable X-ray images, which was done to facilitate subsequent image analysis. The images were then analyzed and annotated by two expert doctors in order to be used to train the AI network. In addition, in order to reduce labeling errors, the dataset was also examined by the third expert doctor. The picture data of this dataset are different in size, and it is necessary to standardize the picture. The X-ray image is first scanned with an X-ray beam on a part of the body, and the detector receives the X-ray light through the part and converts it into visible light, then the photoelectric converter converts it into an electrical signal, and then the analog/digital converter converts it into a digital signal, and finally the digital signal is input into a computer for preprocessing. Due to the particularity of X-ray image, it is necessary to preprocess the image. The dataset contains 5863 X-ray images (all in JPEG format). Before training the four networks used in this paper, the dataset is divided into three parts. These three parts are the training set, the test set, and the verification set. Each category contains pneumonia pictures and normal pictures, as shown in Figure 11.

In this experiment, the dataset is divided into training set (TRAINSET), validation set (VALSET), and test set (TESTSET) according to the ratio of 7:1:2, and the division is random. The specific numbers are shown in Table 1, assigned to 4103 images for the training set, 585 images for the validation set, and 1175 images for the test set. There are three times more pneumonia images in the dataset than normal lung images, resulting in an imbalance in the dataset categories. This study uses 10-fold cross-validation to ensure the validity of the experiment and to avoid the chance of the effect of the convolutional neural networks involved in the experiment. In each experiment, 10% of the label data were randomly selected for verification, and 20% of the label data were tested. The average results of 10 experiments were taken as the final classification results.

Due to the lack of training samples, it is prone to overfitting in the training of convolutional neural networks. That is, when the convolutional neural network is trained, the model performs well on the dataset, while the performance of the test model on other data is very poor. The fundamental reason is that the dataset is too small to train the network by mistake regarding distorted regions or noises in the image as features to be classified rather than as useless information. The overfitting phenomenon caused by too small dataset can be alleviated by data enhancement method. Academia has proposed many data enhancement methods. For example, image clipping, flip, translation, adding noise, rotation, scaling, and other random operations can alleviate the overfitting phenomenon. Recent studies have shown that occlusion, paired samples, and other new data enhancement methods can improve the model effect in image classification and segmentation tasks. Figure 12 below is an example of data enhancement.

In order to overcome the imbalance of image datasets, an image enhancement module is designed. By not changing the conversion of the image class, the number of image files in the dataset is artificially increased. The transformation steps for each image are as follows:

(1) Format conversion. Change the image format to PNG.

(2) Rotation. Rotate the image to 45 degrees with a probability of 0.5.

(3) Flip. Inverts the row pixels in the image array with a probability of 0.5.

(4) Random brightness. When the brightness value is greater than 1, take any value in the set [1.1, 1.3] to improve the brightness of the image.

3.3. Performance Evaluation of Neural Networks. In order to evaluate the advantages and disadvantages of classifiers in deep learning, we set four basic concepts: positive, false, negative, and true, as shown in Table 2.

There are various evaluation metrics, and three common ones are described below: accuracy, precision, and recall.

Accuracy represents the proportion of all correctly predicted samples in the total sample size, and its calculation formula is as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}. \tag{16}$$

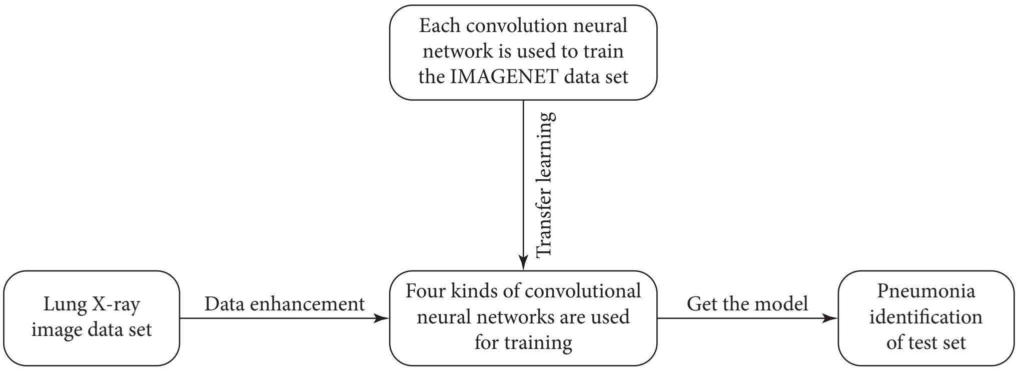

Each convolution neural network is used to train the IMAGENET data set
Transfer learning
Lung X-ray image data set
Data enhancement
Four kinds of convolutional neural networks are used for training
Get the model
Pneumonia identification of test set

Figure 10: System framework.

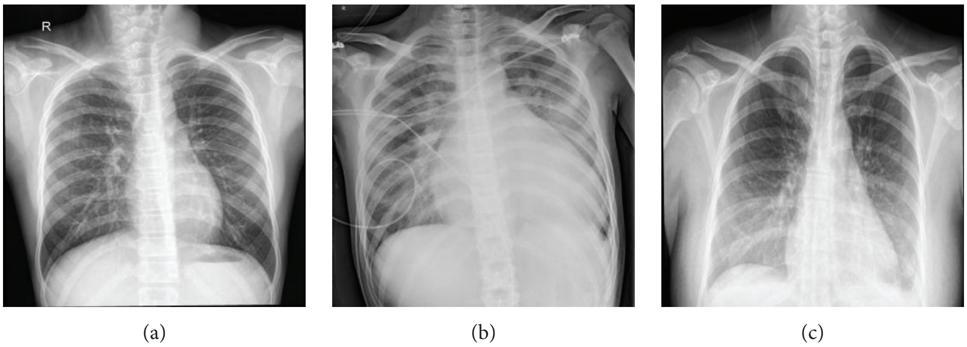


Figure 11: Picture of normal lung, picture of viral pneumonia, and picture of bacterial pneumonia. (a) normal, (b) viral pneumonia, and (c) bacterial pneumonia.

Table 1: Pneumonia data segmentation of chest X-ray images.

Chest X-ray images (pneumonia)				
	VALSET	VALSET	TESTSET	Total
Viral pneumonia	1345	86	148	1579
Bacterial pneumonia	1733	353	734	2820
Normal	1025	146	293	1464
Total	4103	585	1175	5863

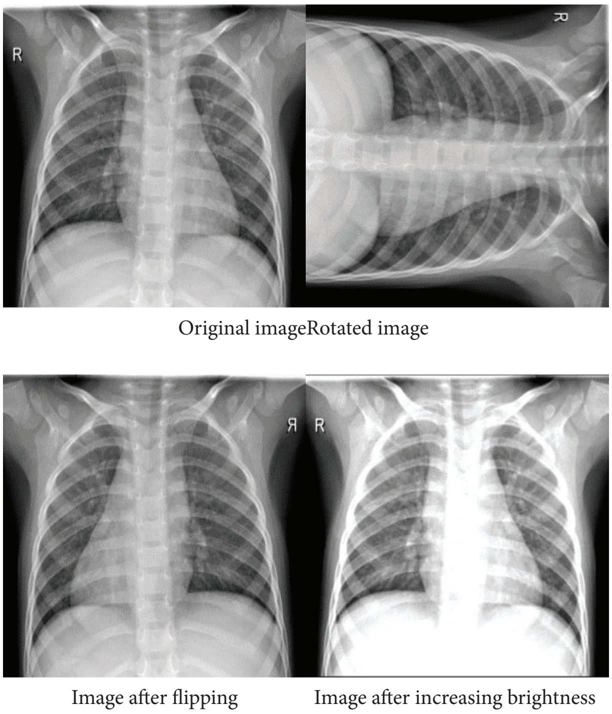
Precision represents the proportion of all true positive (TP) predicted examples to the total positive examples (TP + FP), and the calculation formula is as follows:

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{17}$$

Recall refers to the proportion of all true positive examples (TP + FN) that are determined to be positive examples. It represents the proportion of positive example data that can be identified. The calculation formula is as follows:

$$\text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{18}$$

According to the characteristics of the experimental image data distribution, this paper uses ACC value (accuracy) as the evaluation standard of neural network classification performance.



Original imageRotated image

Image after flipping Image after increasing brightness

Figure 12: Data augmentation.

Table 2: Classifier performance evaluation table.

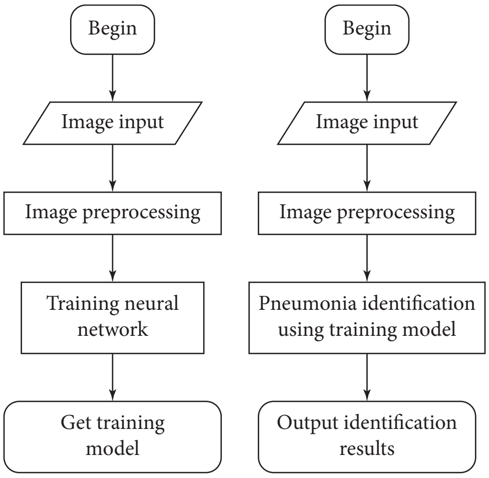
		Positive P	Negative N	Total
The classifier predicts	Positive of prediction, P^*	True positive, TP	False positive, FP	Total positive of prediction, $P^\wedge$
the output	Negative of prediction, N^*	False negative, FN	True negative, TN	Total negative of prediction, $N^\wedge$
Total		Total true positive, P	Total true negative, N	

3.4. Experimental Analysis of Automatic Recognition of Pneumonia Based on DWA Network. In order to verify the actual effect of the convolutional neural network algorithm DWA designed in this paper, four neural network structures (ResNet50, VGG16, DenseNet201, and DWA) are trained on the same platform and condition. The process of the experiment is shown in Figure 13. It can be seen from Figure 13 that the whole experimental process is divided into two parts. The first part of the training sample trains the convolutional neural network until the training requirements are met. The second part uses the training weights obtained in the first part to initialize the convolutional neural network and uses the convolutional neural network to identify new pneumonia as shown in Figure 14. The DWA model identifies statistical data in Table 3.

Data preprocessing, optimizer learning rate dynamic change, padding, and pooling were used in the training process. In this study, 10 crossovers are used to ensure the effectiveness of the experimental convolutional neural network algorithm. To avoid chance in the experimental results, 10% of the overall labeled data were randomly selected for validation and 20% of the labeled data were selected for testing each time; the average result of 10 predictions was taken as the final classification result, and the experiments were conducted 10 times. The experiments were conducted so many times to evaluate and verify the effectiveness of the network model. The performance of the model was improved by tuning the parameters. The results were different each time, and each experiment used the same image enhancement module to overcome the imbalance of the image dataset.

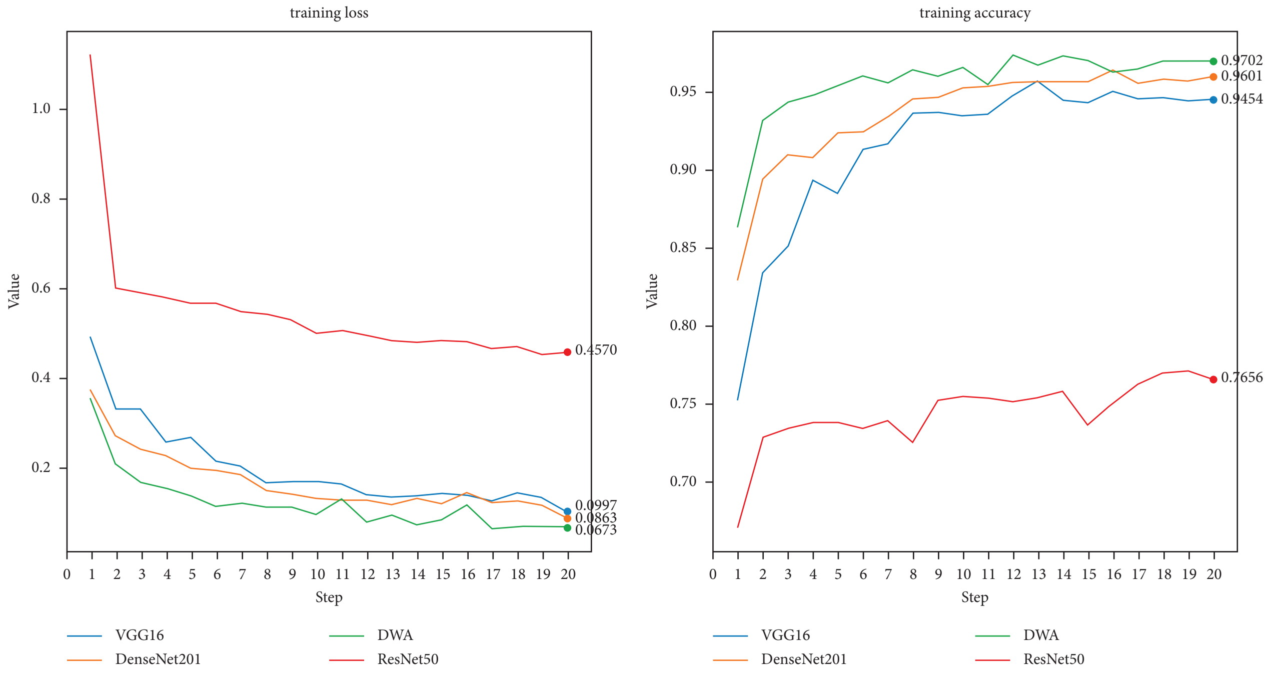
3.5. Analysis of Experimental Results. The ResNet50, DenseNet201, and VGG16 network models were compared with DWA. In the experiment, when the deep convolution and attention neural network (DWA) was trained, the training accuracy reached 97.5%. Since the model is easy to overfit, the verification accuracy is 79%, and the test dataset has 1175 X-ray images, and the test accuracy is 96.1%. The above data show that the deep separable convolutional neural network algorithm based on attention mechanism has higher accuracy in small sample dataset. The experimental results show the effectiveness of attention mechanism and the reliability of deep separable convolutional neural network algorithm.

In the field of pneumonia diagnosis, in the process of innovative research on the identification of medical X-ray images by deep learning algorithm, four convolutional neural networks were compared and relevant conclusions were obtained. The VGG model uses a combination of several small filter convolution layers to replace a large filter convolution layer, and its structure is very simple. Moreover, since the convolution kernel of VGGG focuses on expanding the number of channels and pooling focuses on narrowing the width and height, the model architecture is deeper and wider. At the same time, the increase of calculation amount slows down, and the depth of layers makes the feature map wider. The network can solve 1000 image classification and location problems, but when the depth of the network in VGG reaches a certain degree, the overall performance of the deep network is not as good as the shallow network, that is, degradation problem. ResNet can well solve this problem, because each convolution will waste some information, such as the randomness of convolution kernel parameters (blindness) and the inhibition of activation function. At this time, the shortcut in ResNet is equivalent to take the previously processed information directly to the present processing, which has played a reduction effect. Moreover, ResNet can make the feedforward/feedback propagation algorithm go smoothly with simpler structure, and the increase of identity mapping will not reduce the performance of the network. But ResNet also has some shortcomings, for example, some layers being selectively discarded and information blocking. Compared with ResNet, DenseNet adopts concatenate structure, which can effectively save parameters and reduce computation, thereby saving bandwidth and reducing storage overhead. At the same time, DenseNet algorithm has very strong antioverfitting performance. Compared with the general neural network classifier which directly depends on the characteristics of the last layer of the network (the highest complexity), DenseNet can comprehensively utilize the characteristics of low complexity in the shallow layer, so it is easier to obtain a smooth decision function with better generalization performance. However, DenseNet is very memory-consuming in training. The current deep learning framework does not support the dense connection of DenseNet, so it can only use the repeated Concatenation operation to stitch the output of the previous layer together with the output of the current layer and then transmit it to the next layer. In contrast, the DWA algorithm model can also reduce the number of network parameters and computation and reduce the memory occupation of each parameter by quantifying the parameters. It is much higher than ordinary convolution in terms of parameter number and computational efficiency. If necessary, DWA can greatly improve computational efficiency by sacrificing a small amount of accuracy. In this pneumonia automatic identification algorithm test, DWA is the channel that pays more attention to the most information, while suppressing the unimportant channel characteristics to obtain the optimization of the algorithm. It is the deep separable convolutional neural network, and it is the effectiveness of attention mechanism and the reliability.



(a) (b)

Figure 13: Flow chart of experimental process.



(a) (b)

Figure 14: Training and test results on four network pneumonia datasets. (a) Diagram of loss change process of model in test set. (b) Diagram of changing process of accuracy of the model in the test set.

Table 3: DWA model identifies statistical tables.

Pneumonia type	Total number of tests	Wrong number	Correct number	Accuracy (%)
Viral pneumonia	148	7	141	95.3
Bacterial pneumonia	734	25	709	97
Total recognition rate		96%		

In the process of research, many areas worthy of optimization and improvement are still found. First of all, in this study, medical images in the imaging, due to some reasons, result in changes in the gray value of the image, noise pollution, and details loss, which will affect the image quality, thereby reducing the accuracy of image recognition. Therefore, it is necessary to further study the medical image preprocessing methods. Secondly, there is still much room for improvement in prediction accuracy. Firstly, we can try to use GAN network algorithm. Through the preliminary analysis of the characteristics of GAN network algorithm, it is found that GAN network algorithm can effectively compensate for the lack of datasets. This paper believes that better prediction results can be obtained by improving the generative adversarial network algorithm to test the dataset (such as CycleGAN) so as to improve the accuracy. Secondly, the convolution neural network can be optimized by weak supervision method to train the dataset, which can reduce the dependence on the dataset.

4. Conclusion

In view of the various characteristics of medical image dataset, such as small data scale and unbalanced class distribution, data preprocessing before the experiment (through data enhancement algorithm) and optimization methods in convolution neural network training in the experiment (such as dropout method and automatic change of learning rate of the optimizer) are used to solve these problems. In this experiment, four convolutional neural network structures (ResNet50, VGG16, DenseNet201, and DWA) are analyzed, and experiments are carried out on the same platform and environment. The accuracy of each network on the test set is verified to verify that the proposed deep separable convolutional neural network algorithm based on attention mechanism has higher accuracy in small sample datasets. The effectiveness of attention mechanism and the reliability of deep separable convolutional neural network algorithm are proved.

Over the years, artificial intelligence has been widely used in various fields, especially in the field of biomedical images, and has achieved remarkable results. However, through the investigation of related fields, it is still found that most of the problems in the biomedical field have not been effectively applied or solved. It is artificial intelligence algorithms represented by deep learning in view. This is the satisfactory progress. Deep learning is widely expected to assist doctors in clinical diagnosis. At the same time, in the biomedical field, the industry needs to do more research to solve the interpretability of deep learning networks and how to best maintain the stability of deep learning model prediction. There are still some problems, such as a large amount of data, lack of annotation, and personal privacy information restrictions. In the future, we will effectively improve these problems and make the biomedical field develop faster.

Data Availability

All the data and calculations are in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Xinjiang Uygur Autonomous Region Innovation Environment (Talent, Base) Construction Foundation (Xinjiang NSFC Program Foundation 2020D01A132), Research and implementation of horizontal inversion optimization interpretation method, Hubei Science and Technology Demonstration Foundation (2019ZYYD016), Cooperative Innovation Center of Unconventional Oil and Gas, Yangtze University (Ministry of Education & Hubei Province) (No. UOG2020-10), Logging Simulation of Gas Water Two-Phase Production in Shale Gas Horizontal Wells (2020CB21-25), and Digital core graphic virtual simulation teaching system based on Alibaba cloud.

References

[1] D. Ran, C. Yin, and X. Jia, "Research progress on multi-angle synthetic aperture radar imaging technology," *Journal of Equipment Academy*, vol. 75, no. 3, pp. 57–63, 2016.

[2] D. G. Xie and J. P. Liu, "Research on non-scanning laser radar imaging technology based on compressive sensing," *Electro-Optic Technology Application*, vol. 67, no. 1, pp. 51–59, 2018.

[3] M. Migliaccio, F. Nunziata, C. E. Brown et al., "Polarimetric synthetic aperture radar utilized to track oil spills," *Eos Transactions American Geophysical Union*, vol. 93, no. 16, pp. 161-162, 2013.

[4] M. Similä, M. Mäkynen, and I. Heiler, "Comparison between C band synthetic aperture radar and 3-D laser scanner statistics for the Baltic Sea ice," *Journal of Geophysical Research Oceans*, vol. 115, no. C10, pp. 15–23, 2010.

[5] X. L. Wang and C. X. Chen, "Image fusion for synthetic aperture radar and multispectral images based on sub-band-modulated non-subsampled contourlet transform and pulse coupled neural network methods," *The Imaging Science Journal*, vol. 64, no. 2, pp. 87–93, 2016.

[6] K. C. Jezek, "Glaciological properties of the Antarctic ice sheet from RADARSAT-1 synthetic aperture radar imagery," *Annals of Glaciology*, vol. 29, no. 1, pp. 286–290, 2017.

[7] L. M. Novak and C. M. Netishen, "Polarimetric synthetic aperture radar imaging," *International Journal of Imaging Systems & Technology*, vol. 4, no. 4, pp. 306–318, 2010.

[8] D. L. Schuler, J. Lee, T. L. Ainsworth, and M. R. Grunes, "Terrain topography measurement using multipass polarimetric synthetic aperture radar data," *Radio Science*, vol. 35, no. 3, pp. 813–832, 2016.

[9] J. Bai, B. Liu, L. Wang, and L. Jiao, "PolSAR image compression based on online sparse K-SVD dictionary learning," *Multimedia Tools & Applications*, vol. 76, no. 11, pp. 1–12, 2017.

[10] R. Sharma and R. K. Panigrahi, "Stokes based sigma filter for despeckling of compact PolSAR data," *IET Radar, Sonar & Navigation*, vol. 12, no. 4, pp. 475–483, 2018.

[11] J.-S. Lee, J.-H. Wen, T. L. Ainsworth, K.-S. Chen, and A. J. Chen, "Improved sigma filter for speckle filtering of SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 202–213, 2009.

[12] J. S. Lee, T. L. Ainsworth, Y. Wang, and K.-S. Chen, "Polarimetric SAR speckle filtering and the extended sigma filter," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 53, no. 3, pp. 1150–1160, 2014.

[13] E. Karami, S. Prasad, and M. Shehata, "Image matching using SIFT, SURF, BRIEF and ORB. Performance comparison for distorted images," pp. 28–37, 2017, https://arxiv.org/abs/1710.02726.

[14] Y. Chen and L. Shang, "Improved SIFT image registration algorithm on characteristic statistical distributions and consistency constraint," *Optik*, vol. 127, no. 2, pp. 900–911, 2016.

[15] A. N. Ting, H. E. Yi-Min, and Z. Y. Zhang, "An improved bidirectional SIFT feature matching algorithm," *Computer Engineering & Science*, vol. 32, no. 3, 2016.

[16] C. Song, H. Deng, H. Gao, H. Zhang, and W. Zuo, "Bayesian non-parametric gradient histogram estimation for texture-enhanced image deblurring," *Neurocomputing*, vol. 197, no. C, pp. 95–112, 2016.

[17] C. Sbarufatti, M. Corbetta, M. Giglio, and F. Cadini, "Adaptive prognosis of lithium-ion batteries based on the combination of particle filters and radial basis function neural networks," *Journal of Power Sources*, vol. 344, pp. 128–140, 2017.

[18] E. Larsson, V. Shcherbakov, and A. Heryudono, "A least squares radial basis function partition of unity method for solving PDEs," *SIAM Journal on Scientific Computing*, vol. 39, no. 6, pp. 45–62, 2017.

[19] X. Guo, S. Shao, N. Ansari, and A. Khreishah, "Indoor localization using visible light via fusion of multiple classifiers," *IEEE Photonics Journal*, vol. 9, no. 6, pp. 1–10, 2017.

[20] G. M. Foody, L. See, S. Fritz et al., "Assessing the accuracy of volunteered geographic information arising from multiple contributors to an Internet based collaborative project," *Transactions in GIS*, vol. 17, no. 6, pp. 847–860, 2013.

[21] D. Yenigün, G. Ertan, and M. Siciliano, "Omission and commission errors in network cognition and network estimation using ROC curve," *Social Networks*, vol. 50, pp. 26–34, 2017.

[22] S. C. Liu, Z. M. Wen, and T. Yu, "Influence of different topographic correction methods on the remote sensing extraction of *Robinia pseudoacacia* distribution," *Journal of Beijing Forestry University*, vol. 39, no. 5, pp. 25–33, 2017.

[23] S. Sanders, D. Flaws, M. Than, J. W. Pickering, J. Doust, and P. Glasziou, "Simplification of a scoring system maintained overall accuracy but decreased the proportion classified as low risk," *Journal of Clinical Epidemiology*, vol. 69, pp. 32–39, 2016.

[24] C. Niu, W. K. Wong, and Q. Xu, "Kappa ratios and (higher-order) stochastic dominance," *Risk Management*, vol. 19, no. 4, pp. 1–9, 2017.