PCXRNet: Condense attention block and Multiconvolution spatial attention block for Pneumonia Chest X-Ray detection

Yibo Feng, Xu Yang, Dawei Qiu, Huan Zhang, Dejian Wei, Jing Liu

Abstract—Pneumonia is one of the leading causes of worldwide child deaths. Many recognition approaches based on convolutional neural networks have been proposed for children's chest X-ray images in recent years. However, few of them make good use of the potential inter- and intra-relationship of feature maps. Considering the challenge mentioned above, this paper proposes an attention-based convolutional neural network, named PCXRNet, for pneumonia chest X-ray diagnosis. To utilize the information on the channels of feature maps, we add a novel condense attention module (CDSE), containing two steps: condense step and squeeze-excitation step. Unlike traditional channel attention modules, CDSE first downsamples the feature map channel by channel to condense the information before the squeeze-excitation step, in which channel weights are calculated. In order to make the model pay more attention to informative spatial parts in every feature map, we propose a multi-convolution spatial attention module (MCSA). It reduces parameters and introducing more nonlinearity. The CDSE and MCSA can work in series to dig up useful information on and between feature maps complementarily. We conduct extensive experiments on ChestXray2017, and compared with existing methods, PCXRNet reaches remarkable performance with the accuracy of 97.436%, recall of 96.838%, precision of 97.705%, and F1-score of 97.241%. We also use the other two pneumonia datasets COVID-19 and Tuberculosis, where PCXRNet achieves state-of-the-art results. A gradient-weighted class activation mapping algorithm is applied to further demonstrate the effectiveness of our method.

Index Terms—deep learning, pneumonia detection, attention mechanism, chest X-ray images, convolutional neural network.

I. INTRODUCTION

THE lung is typically one of the most crucial organs in the human respiratory system. Pneumonia is often caused by bacteria or viruses with cough, fever, and difficulty breathing. According to the United Nations Children's Fund, pneumonia is one of the leading causes of worldwide child deaths in many infectious diseases.

This work was funded by National Natural Science Foundation of China (8143708, 81973981), Natural Science Foundation of Shandong Province (ZR2020MH360), Teaching cases library construction project for graduate students with professional degrees of Shandong Province (SDYAL20049), Traditional Chinese Medicine Science and Technology Development Program of Shandong Province (2019-0056);

Y.B. Feng, D.W. Qiu, H. Zhang, D.J. Wei and J. Liu are with the College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan 250355, China (Corresponding author: J. Liu Email: liuj_jn@163.com). X. Yang are with the Megvii, Beijing 100000. Y.B. Feng and X. Yang contribute equally.

Fig. 1. Examples of Chest X-rays in ChestXray2017. The normal, viral pneumonia and bacterial pneumonia are on the left, middle and right, respectively.

In 2016, more than 880,000 children died of pneumonia [1], and the number of child deaths caused by it exceeds malaria, AIDS, and yellow fever combined. The incidence of childhood pneumonia shows a significant difference between developed and underdeveloped countries [2], [3]. Childhood pneumonia may cause respiratory failure and easily lead to life-threatening if not treated quickly enough. However, most patients can be cured through early detection and treatment [4], [5].

Chest X-ray images can detect lung lesions, which can help doctors judge the severity of lung infections [6]. Relying solely on radiologists for diagnosis of childhood pneumonia has the following problems. Firstly, the proportion of doctors and patients is seriously imbalanced in low-income countries with severe children's pneumonia. Secondly, the diagnostic accuracy is highly influenced by the radiologists' experience and carefulness. Finally, the increasing number of patients infected brings a heavy burden for radiologists to finish the diagnostic process, and fatigue may impact diagnostic accuracy [7]. Therefore, it is indispensable to use computer-aided diagnosis technology to help radiologists increase the detection speed and relieve the burden.

Classical Machine Learning (ML) methods cannot extract the high-dimensional features of the medical image well and have poor accuracy and generalization ability [8], [9]. Deep learning, particularly with convolutional neural networks (CNNs), has achieved great success in computer vision. Compared to the traditional image processing algorithms, CNNs use relatively few pre-processing methods and can be trained in an end-to-end way. Therefore, CNNs are usually viewed as a general feature extractor with no need for prior knowledge and human intervention. CNNs have achieved remarkable performance on many tasks, such as classification [10], object detection [11] and semantic segmentation [12]. Deep learning is also used to detect lung diseases with the release of medical imaging datasets. Lessmann et al. [13] proposed a two-step method to detect the lung disease in low-dose chest CT. The first CNN identifies potential lesion

locations using dilated convolution and the second CNN identifies true calcifications. Aledhari et al. [14] explored the effects of three basic CNNs, VGG16, Inception v3 and ResNet-50, on pneumonia detection. Fine-tuned VGG16 demonstrates the most test accuracy of 75%.

Although computer-aided diagnosis technology has made some progress in distinguishing childhood pneumonia, it is still challenging to apply deep learning technology to clinical diagnosis due to the complexity of the pulmonary lesions. First, the potential relationship between feature maps is ignored by traditional CNNs algorithms, which do not suppress the useless information in the feature maps. Second, unlike adults, children can receive lower radiation doses, and the treatment compliance is poorer [15], which makes the lesions of the chest X-ray image blurred and deformed. Third, previous CNN-based pneumonia detection methods have poor generalization ability due to the overfitting problems.

Considering the above-mentioned problems, we propose an attention-based convolutional neural network for the rapid detection of childhood pneumonia. For the first problem, this paper introduces a novel channel attention module, named condense attention module (CDSE), to generate more accurate channel weights of feature maps. Compared with traditional channel attention blocks, CDSE downsamples each channel of the input feature map to the appropriate size using depthwise convolution before the squeeze-excitation block. The residual structure is adopted by CDSE to enhance the representation ability of the network. For the second problem, we propose a multi-convolution spatial attention module (MCSA), a lightweight module using three small convolutions instead of a large convolution. MCSA has smaller parameters and more nonlinearity, suppressing overfitting and enhance the ability to learn spatial information from feature maps. We embed the two proposed attention modules into a baseline network for pneumonia chest X-ray recognition, called PCXRNet, which has better generalization performance and can more effectively identify the lesions in chest X-ray images. As is shown in Fig 1, the chest X-ray examples include normal, viral, and bacterial pneumonia. The main contributions of this paper include the following.

(1) A novel channel attention module (CDSE) is proposed to explore appropriate channel weights of input feature maps. CDSE firstly condenses information using depthwise convolution to gradually downsample feature map of baseline network's different stages to a fixed size with dynamic adaptive strategy and then feeds condensed information to squeeze-excitation block. Experimental results show that embedding CDSE into baseline can enhance the generalization ability and bring a significant improvement to pneumonia recognition

(2) A lightweight multi-convolution spatial attention module (MCSA) is proposed to local lesion on feature maps while suppresses useless spatial information. CDSE and MCSA are two complementary attention modules. The extensive ablation

experimental results show that integrating the two attention modules in a sequential manner makes the network more stable and suppresses the transmission of useless information.

(3) Additional experiments are conducted on a COVID-19 dataset with 18,479 X-ray images and a tuberculosis dataset with 7000 images in order to verify the effectiveness of the proposed PCXRNet. As a result, the network not only performs best in childhood pneumonia but also achieves state-of-the-art results on the COVID-19 dataset and tuberculosis dataset.

II. RELATER WORKS

In this section, we present the related works on pneumonia recognition and attention mechanisms in deep learning.

A. Automatic pneumonia diagnosis in deep learning

Deep learning applications in medical diagnosis continue to be catalyzed and expanded by the increasing release of medical image datasets [16], [17], [18], [19], [20]. Trained with a large amount of data, CNNs can achieve the same accuracy as radiologists but have a higher diagnostic efficiency [21]. Deep learning algorithms can solve the radiologist shortage for areas with insufficient medical resources. Kermany et al. [22] collected and labeled 5856 chest X-ray images, including normal, viral, and bacterial pneumonia. Inception V3 pretrained on the ImageNet dataset is used to recognize pneumonia. Hu et al. [23] proposed the multi-kernel depthwise convolution (MD-Conv), which contains different filter sizes in one convolution layer. Experiments on the pneumonia dataset are conducted to verify the generalization ability. The proposed network achieved an accuracy of 93.4%. Masud et al. [24] identified the presence of pneumonia through analyzing chest X-ray images. The augmentation technique was used to balance the number of pneumonia radiographs. The chest X-ray images resized to 256×256 were used to extract features using a statistical and a deep learning algorithm. These features are combined and sent to the random forest for classification. Previous works have proved that deep learning networks can effectively diagnose childhood pneumonia in chest X-ray images. However, they do not make full use of the information of feature maps and hence cannot be promoted in clinical diagnosis.

Corona Virus Disease 2019 (COVID-19) has rapidly become a global epidemic because SARS-CoV-2 is highly contagious. It is necessary to apply deep learning technology to diagnose COVID-19 from chest radiographs. Wang et al. [25] used 1136 images, including 723 COVID-19 positives, to train the deep learning model. The dataset was collected from five hospitals and they organized several senior physicians to choose the required data. The proposed system first extracted the lung lesion regions using 3D U-Net++, and then the segmentation results were sent into ResNet50 to get classification results. Ghoshal et al. [26] proposed a multi-classification network for diagnosing COVID-19, pneumonia, and lung cancer from a joint dataset including chest x-ray and CT images.

inter-channel relationship of features by learning the channel-wise attention to generate weighted feature maps $F_{CDSE_out} \in R^{H \times W \times C}$. Specifically, unlike traditional channel attention mechanism, CDSE is designed with two components, i.e., Channel Feature Extract (CFE) block and mSE block, where CFE first applies depthwise convolution to condense information channel by channel and then mSE computes each channel weight with global average pooling and 1×1 convolution layers to select more useful features and suppress less informative features. Next, given feature maps F_{CDSE_out} , multiconvolution spatial attention module (MCSA) produces $F_{MCSA_out} \in R^{H \times W \times C}$, by learning inter-spatial relationship of features and multiplying extracted spatial-wise attention weight to the input indicating the importance of each spatial position.

We aggregate the two attention modules, CDSE and MCSA, in a sequential order to complementarily exploit the inter-channel and inter-spatial relationship of features, bringing a significant performance improvement in recognition of chest X-ray images, which is supported by experimental results.

B. Condense channel attention module

(1) Structure: Constructed of several CFE blocks and one mSE block, CDSE module is designed for selecting useful information and suppressing useless information in chest X-ray images more effectively to improve the classification performance of the network. CFE blocks apply depthwise convolution, IEBN [40] and Mish function [41] to condense information from feature maps for every channel. We replace the ReLU activation function in the original SE block with the Mish function to propose the mSE block, considering the better capability of Mish such as transmission of gradients, representation ability and generalization. The mSE block squeezes CFE's output to produce the excitation of channel-wise attention weights. Note that we adopt the dynamic adaptive strategy to determine the number of CFE blocks, making the size of the smallest output of CFE blocks 7×7 in every residual block for four stages, which is consistent with the smallest output of the last stage without CDSE module. Therefore, 3, 2, 1 CFE blocks are used in the first three stages, and none for the last stage.

Given an intermediate feature map $F_{residual_out}$ from residual connection as input, CFE block product output $F_{CFE} \in R^{H \times W \times C}$ as follows:

$$F_{CFE} = \text{Mish}(\text{CFE_res}(F_{ro}) + \text{CFE_shortcut}(F_{ro})) \quad (1)$$

where CFE_res and CFE_shortcut represent the residual connection and the skip connection of the CFE block, respectively. F_{ro} is $F_{residual_out}$. CFE_res uses two depthwise convolution with the kernel size of 3×3 to extract features and the stride of the former one is 2. CFE block adopts IEBN [40] to normalize the output of convolution, which can refine the noise produced by traditional BN, and use Mish[41] as activation function for introducing nonlinearity. Compared with ReLU, Mish has the smoother gradient and relieve the dying ReLU problem, so it provides better transmission of gradient when training. The formula of Mish is shown as:

$$\text{Mish}(x) = x \times \tanh(\ln(1 + e^x)) \quad (2)$$

With the input $F_{mSE_in} \in R^{7 \times 7 \times C}$ obtained by stacking several CFE blocks, mSE block first squeezes feature maps to feature vectors with the size of $1 \times 1 \times C$.

$$f = \frac{1}{H \times W} \sum_{j=1}^H \sum_{k=1}^K u_i(j, k) \quad (3)$$

where H and W denote the height and width of the feature maps, respectively and $u_i(i = 1, 2, \dots, C)$ represents the i -th channel of the feature map.

Then, mSE uses convolution to produce the channel-wise attention weights. Note that we also set reduction ratio to vary the number of channels between two convolution layers and choose the best ratio through experiments. Finally, the output of CDSE module F_{CDSE_out} is calculated as follows:

$$F_{CDSE_out} = \sigma(\text{conv2}(\text{GAP}(F_{mSE_in}))) \otimes F_{ro} \quad (4)$$

where GAP denotes global average pooling and conv2 indicates two 1×1 convolutional layers. σ is Sigmoid activation function. $\otimes$ denotes element-wise multiplication.

(2) Analysis

First, we analyze two methods to condense the information of feature maps before squeeze operation. The first is to use the fixed number of CFE blocks. The other is to adopt the proposed dynamic adaptive strategy to determine the number of CFE blocks in each residual block. The second method condense feature maps to a fixed size of 7×7 for every residual block, which is then fed into mSE block. Compared with the former, dynamic adaptive strategy reduces extra parameters. For example, the first method adopts the fixed number of 2 CFE blocks in each CDSE module introducing additional 68 convolution layers, yet the second method only adds 46 additional convolution layers, which reduces nearly one in three. In addition, using the adaptive strategy is beneficial for extracting channel-wise weights, for making the size of input for mSE block in different stage the same.

Second, we adopt a new operation combination of Conv-IEBN-Mish to replace Conv-BN-ReLU in the CDSE module, which has a better capability of generalization and improves the classification performance significantly with the small additional computational cost.

Third, we improve the traditional SE module using Mish as the activation function. CDSE module aggregates several stacked CFE blocks and the mSE block to extract global feature maps and study the channel-wise attention weights, which select important information related to the disease and suppress useless information of feature maps.

C. Multi-convolution spatial attention block

(1) Structure: Multi-convolution spatial attention module

Fig. 3. The overall architecture of multiconvolution spatial attention block (MCSA)

(MCSA) is designed for extracting spatial information from feature maps, of which the structure is shown in Fig.3. Inspired by CBAM [42], MCSA adopts two-branch spatial attention mechanism, instead of using 7×7 convolution, three conservative 3×3 convolutions are stacked to introduce more non-linearity with the same receptive fields but fewer parameters, which can further boost the model accuracy. Also, we use Mish as activation function.

Given input feature map $F_{MCSA_in} \in R^{H \times W \times C}$, MCSA uses max pooling and average pooling to extract spatial information of feature maps, and then outputs feature maps $F_{max} \in R^{H \times W \times 1}$, $F_{avg} \in R^{H \times W \times 1}$, respectively. The process can be described as:

$$F_{max} = \max(u_i(j, k)) \quad (5)$$

$$F_{avg} = \frac{1}{C} \sum_{i=1}^C u_i(j, k) \quad (6)$$

where $j \in \{1, 2, \dots, H\}$, $k \in \{1, 2, \dots, W\}$, $u_i(1, 2, \dots, C)$ is i -th channel of input feature maps. Max pooling and average pooling gather spatial information from different aspects. Then, the output before convolution $F_{cat} \in R^{H \times W \times 2}$ is obtained as follows:

$$F_{cat} = \text{Concat}(F_{max}, F_{avg}) \quad (7)$$

After concatenating average-pooled and max-pooled features, MCSA uses convolution layers to output the spatial mask, which finally converts the input feature map F_{MCSA_in} to $F_{MCSA_out} \in R^{H \times W \times C}$ by the following formula:

$$F_{MCSA_out} = \sigma(\text{Conv}(\text{ConvM2}(F_{cat}))) \otimes F_{MCSA_in} \quad (8)$$

where σ is the sigmoid function and $\otimes$ denotes element-wise multiplication. Conv and ConvM2 represents two layers of 3×3 convolution without and with Mish function, respectively.

(2) Analysis

MCSA adopts two-branch structure to learn spatial-wise attention from feature maps. It uses three consecutive convolutions with smaller kernel size of 3×3 to replace original 7×7 convolution, in order to reduce parameters under the same receptive field. For the use of Mish function, our

experimental result shows that it exactly works better than ReLU function. CDSE and MCSA are two different attention modules, where CDSE focuses on extracting channel-wise attention weights, whereas MCSA mainly extract spatial-wise attention weights to capture disease specific information from feature maps. Aggregating two attention modules in a sequential order brings a significant performance improvement for chest X-ray classification

IV. EXPERIMENT

In this section, we first introduce the chest X-ray datasets on pneumonia and common evaluation metrics, then present the experimental results on three pneumonia datasets, and interprets the effectiveness of the proposed attention mechanism through visualization.

A. Datasets

ChestXRay2017: This dataset was selected from retrospective study of pediatric patients aged from one to five years old in Guangzhou Women and Children's Medical Center, Guangzhou [22]. This dataset contains a total of 5856 chest X-ray images. The training set contains 3883 pneumonia images (2538 bacterial images and 1345 viral images) and 1349 normal images, and the test set contains 390 pneumonia images (242 bacterial and 148 viral) and 234 normal images.

COVID-19 Radiography Database V4: This dataset was created by a team of researchers from Qatar University, the University of Dhaka, etc. [43]. It is currently one of the largest datasets of chest X-ray images of COVID-19. We used 80% of the chest X-ray images for training and 20% for testing to split data. And 20 % of images in the training dataset were utilized as a subset for validation. Table 1 shows the details of the number of training, validation and test used in this paper.

TABLE 1
DETAILS OF THE COVID-19 DATASET

	COVID-19	Lung Opacity	Normal	Total
Train	2314	3847	5664	11825
Validation	578	962	1416	2956
Test	724	1203	1771	3698
Total	3616	6012	8851	18479

Tuberculosis Chest X-ray Database: This dataset is composed of four public datasets [44], including NLM dataset[45], Belarus dataset[46], NIAID TB dataset[47] and RSNA CXR dataset[48]. This dataset is one of the largest tuberculosis datasets so far. Our experiment uses 80% of images for training, 20% for testing, and 20% of the training set for validation. Table 2 shows the details of the tuberculosis dataset.

TABLE 2
DETAILS OF THE TUBERCULOSIS DATASET

	Tuberculosis	Normal	Total
Train	2240	2240	4480
Validation	560	560	1120
Test	700	700	1400
Total	3500	3500	7000

B. Evaluation criteria

In order to comprehensively measure the effectiveness of the proposed network, we employ accuracy, recall, precision and F1-score as evaluation indicators. The calculation formulas are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

C. Ablation Experiments on ChestXRay2017

(1) Implementation Details

In ablation experiments, the input images of PCXRNet were resized to 224×224 pixels. The initial learning rate was 0.01 and the maximum epoch was 200. At the 15th, 65th and 150th epoch, the learning rate was reduced by factor 10, respectively. The network was trained using the batch size of 16 with SGD optimizer and cross-entropy loss function. Our framework was implemented with PyTorch library and all experiments were conducted on NVIDIA GTX 2080Ti.

(2) The results of ablation experiments

In order to verify the effectiveness of the attention module proposed in this paper, we show the results of ablation experiments on ChestXRay2017 in Table 3. Note that all experiments in this paper have not used any pre-trained models for a fair comparison.

In Table 3, model 1 and model 2 refers to ResNet34 [10] and SE-ResNet34 [36], respectively. Model 3 introduces the CDSE attention module based on ResNet34. The introduction of the proposed channel attention module can improve the accuracy, recall, precision and F1-score by 1.28%, 1.54%, 1.19% and 1.40%, respectively, compared with the baseline.

Model 4 introduces the spatial attention module (SA) of CBAM [42] into the baseline. It first adopts average pooling and max pooling along the channel dimension, concatenates the output, and then generates spatial mask. MCSA module is added into model 5. The introduction of the MCSA can improve the accuracy, recall, precision and F1-score by 0.48%, 0.47%, 0.56% and 0.52%, respectively, compared with model 4. As can be seen, the MCSA module has stronger stability by using 3×3 convolution and introducing more activation layers.

Model 6 introduces the CBAM module into the baseline, which is a lightweight attention module. Model 7 is our proposed method PCXRNet, which integrated the CDSE module and MCSA module in series at the end of each residual block. Results show that the classification performance of PCXRNet is better than CBAM or only using either CDSE module or MCSA module. PCXRNet achieves an overall

accuracy of 97.436%, recall of 96.838%, precision of 97.705%, and F1-score of 97.241%.

It can thus be seen that the CDSE module and MCSA module are complementary. With two modules in combination, our network can present a good classification performance even not using pretraining. Training and testing curves in Fig. 4 demonstrate that the accuracy curve of training increases rapidly from epoch 0 to epoch 40 and then it gradually stabilizes until 200 rounds. The confusion matrix of PCXRNet and other networks is presented in Fig. 5 and Fig. 6, respectively. Observed from the confusion matrix, PCXRNet has the highest accuracy whether in normal images or pneumonia images.

TABLE 3
ABLATION EXPERIMENTS

model	Acc/%	Rec/%	Pre/%	F1/%	ΔAcc/%
baseline [10]	95.513	94.444	96.024	95.135	-
baseline + SE [36]	96.314	95.769	96.349	96.044	↑0.80
baseline + CDSE	96.795	95.983	97.214	96.538	↑1.28
baseline + SA [42]	96.635	96.197	96.608	96.395	↑1.12
baseline + MCSA	97.115	96.667	97.169	96.907	↑1.60
baseline + CBAM [42]	96.795	96.325	96.823	96.563	↑1.28
baseline + CDSE + MCSA	97.436	96.838	97.705	97.241	↑1.92

Fig. 4. Accuracy curve of training set and test set of PCXRNet on ChestXRay2017

Fig. 5. Confusion matrix of PCXRNet on ChestXRay2017

Fig. 6. Confusion matrix of other models on ChestXRay2017

D. Ablation Experiments on ChestXRay2017

In this section, we explore the internal structure of CDSE.

(1) the number of CFE module

There are two ways to reduce the size of the feature map: the first way is to insert the same number of CFE blocks in each residual block; the second is to insert a different number of CFE blocks in each residual block. We designed three experiments to explore a suitable way: the first model named CFE_1 uses one CFE block to downsample feature maps two times in each residual block; the second model named CFE_2 uses two CFE blocks to downsample four times in each residual block; the third model named CFE_stage adopts the dynamic adaptive strategy to determine the number of CFE blocks, making the size of the output of CFE blocks the same as the smallest output of baseline's last stage. Table 4 shows the classification performance of three experiments and ROC curves are shown in Fig. 7(a).

Results show that the model CFE_stage achieves the best classification performance. It has several advantages. First, in the traditional channel attention module, global information was squeezed directly into a channel descriptor regardless of the size of input feature maps. CFE_stage condenses the global information of feature maps before squeezing, which is beneficial for extracting channel-wise weights more accurately. Second, CFE_stage has fewer parameters than CFE_1 and

CFE_2, improving the inference speed and relieving the overfitting problem. As shown in Fig. 7(a), CFE_stage also displays the highest AUC.

TABLE 4
THE RESULTS OF DIFFERENT NUMBER CFE MODULES

	Acc/%	Rec/%	Pre/%	F1/%
CFE_1	96.314	95.598	96.535	96.030
CFE_2	95.673	94.744	96.038	95.322
CFE_stage	96.795	95.983	97.214	96.538

(2) the downsample layer of CFE shortcut

In the shortcut connection of the CFE block, the input feature map is downsampled, and the output is then added to the output of the residual connection. There are many downsample ways: 1×1 convolution, average pooling and max pooling. Different downsample layers will affect the information distribution in the feature map. Table 5 shows the classification performance of different downsample layers and ROC curves are shown in Fig. 7(b). It is obvious that using 1×1 convolution achieves better accuracy.

TABLE 5
THE RESULTS OF DIFFERENT DOWNSAMPLE LAYERS

	Acc/%	Rec/%	Pre/%	F1/%
maxpool	96.314	95.427	96.741	96.015
avgpool	95.994	95.171	96.285	95.676
Conv1 $\times$ 1	96.795	95.983	97.214	96.538

(3) the downsample location of CFE_res

There are two convolution layers in the residual connection of CFE blocks. We conducted a set of comparative experiments to explore where to downsample. As shown in Table 6, downsample in the first convolution layer has a better effect than in the second convolution, improving the accuracy by 0.641%. Also, downsampling in the second layer increases memory usage. ROC curves are shown in Fig. 7(c).

Fig. 7. Comparison of receiver operating characteristic curves on ChestXRay2017 dataset.

TABLE 6
THE RESULTS OF THE DOWNSAMPLE LOCATION

	Acc/%	Rec/%	Pre/%	F1/%
1	96.795	95.983	97.214	96.538
2	96.154	95.838	95.214	96.620

(4) the activation function of CFE module

The activation function is essential for CNNs because it can strengthen the learning ability of the network. ReLU is a common activation function, but it suffers from the dying neuron problem, leading to the loss of gradient information, so it is difficult for the CNNs to learn effective lesion information on chest X-rays. Mish is a self-regularized non-monotonic activation function and preserves a few negative values. Table 7 shows the classification performance for different activation functions. The results in Fig. 7(d) demonstrate that Mish performed better than ReLU.

TABLE 7
THE RESULTS OF DIFFERENT ACTIVATION FUNCTION IN CDSE

	Acc/%	Rec/%	Pre/%	F1/%
ReLU	96.474	95.385	97.205	96.170
Mish	96.795	95.983	97.214	96.538

(5) the normalization layer of CFE module

Normalization layer reduces the dependence on their initial values, which is beneficial for gradient flow. Batch normalization mitigates the internal covariate shift problem and improve generalization capability. IEBN introduces the self-attention mechanism to further improve the generalization, which can refine the noise produced by traditional BN. As

shown in Table 8, IEBN achieves superior classification performance. ROC curves are shown in Fig. 7(e).

TABLE 8
THE RESULTS OF DIFFERENT NORMALIZATION LAYER

	Acc/%	Rec/%	Pre/%	F1/%
BN	96.474	95.897	96.567	96.213
IEBN	96.795	95.983	97.214	96.538

(6) the choice of r in mSE

In this part, we study the important hyperparameter the reduction ratio r in mSE block, which represents the capacity of the channel attention module. Table 9 shows that as r increases from 2 to 32, the network performance enhances then drops. When r is set to 16, the model reaches the highest performance. ROC curves are shown in Fig. 7(f).

The above experiments and discussions finally determined the structure of CDSE. We adopt Conv-IEBN-Mish, a new combination of operations, to replace the traditional Conv-BN-ReLU combination in the CDSE module. This structure generates better gradient flow and extracts channel weights of feature maps more accurately. We adopt a different number of CFE blocks with the dynamic adaptive strategy to

condense the information from feature maps before use mSE block to generate channel descriptors.

TABLE 9
THE RESULTS OF DIFFERENT R

r	Acc/%	Rec/%	Pre/%	F1/%
2	95.994	95.256	96.186	95.684
4	95.994	95.085	96.389	95.668
8	96.635	95.684	97.205	96.354
16	96.795	95.983	97.214	96.538
32	96.154	94.957	96.973	95.813

E. The Structure of MCSA block

In this section, we explore the structure and the location of MCSA.

(1) the activation function in MCSA module

This part discusses the impact of different activation functions on the MCSA module. For experiments of this part, we do not use CDSE module, only place the MCSA module at the end of the residual block. Table X shows the classification performance of different activation functions and ROC curves are shown in Fig. 7(g). The introduction of the Mish can improve the accuracy, recall, precision and F1-score by 0.80%, 0.98%, 0.73% and 0.87%, respectively.

TABLE 10
THE RESULTS OF DIFFERENT ACTIVATION FUNCTION IN MCSA

	Acc/%	Rec/%	Pre/%	F1/%
ReLU	96.314	95.684	96.439	96.037
Mish	97.115	96.667	97.169	96.907

(2) the location of MCSA module

In this part, we mainly discuss the position relationship between MCSA and CDSE. Because MCSA and CDSE focus on extracting effective spatial information and channel information in feature maps, respectively, aggregating two attention modules in a sequential order brings a significant performance improvement for chest X-ray classification. Table 11 shows that the CDSE-first sequential manner can improve the accuracy by 0.641%. This sequential manner can strengthen the learning ability of the network. Results show that CDSE and MCSA are complementary attention modules. ROC curves of different arrangements are shown in Fig. 7(h).

TABLE 11
THE RESULTS OF THE DIFFERENT LOCATION OF MCSA MODULE

	Acc/%	Rec/%	Pre/%	F1/%
CDSE & MCSA	96.795	96.325	96.823	96.563
CDSE + MCSA	97.436	96.838	97.705	97.241

Fig. 8. The visualization results of ResNet34, SE-ResNet34 and PCXRNet on ChestXray2017 dataset. The top row shows original pneumonia images. The second row shows the results without attention. The third row shows the results of SE-ResNet34. The fourth row shows the results of PCXRNet.

F. Visualization of Network

In order to verify the effectiveness of our proposed method, we visualize the output of CNNs by using Gradient-weighted Class Activation Mapping (Grad-CAM) [49]. As shown in Fig. 8, the first row is five pneumonia chest X-ray images from the ChestXray2017 dataset, and the second and third rows are the visualization results of ResNet34, SE-ResNet34 respectively. The visualization results of SE-ResNet34 are better than of ResNet34. The introduction of the SE attention module can suppress useless information in chest X-ray images more effectively than the network without attention, as shown in CXR0, CXR1 and CXR3 in the second row. The visualization results of PCXRNet are shown in the fourth row, which are better than ResNet34 and SE-ResNet34, indicating that PCXRNet can utilize the lesion information in the chest X-ray images more effectively.

G. Comparisons with Other CNNs on ChestXray2017

This part compares the classification results of PCXRNet with other state-of-the-art CNNs on ChestXray2017. Asnaoui et al. [50] preprocess the chest X-ray images with intensity normalization. The pre-trained CNNs are used to recognize pneumonia images. Liang et al. [51] combines residual network and dilated convolution to recognize childhood pneumonia. The proposed method was pre-trained on the ChestX-ray14 dataset to improve classification performance. Chouhan et al. [52] proposed an ensemble model including five CNNs

pre-trained on ImageNet to diagnose pneumonia. Our model PCXRNet does not use any pre-trained model in all experiments but still achieves the best performance in accuracy and F1-score of 97.46% and 97.241%, respectively, clearly seen from Table 12. This demonstrates that our proposed method has higher accuracy and stability. Model 9 has the best recall reaching 99.62% and outperforms our proposed PCXRNet by 2.78%, but its precision of 93.28% is much lower than our method. PCXRNet achieves a precision of 97.705% and outperforms model 9 by 4.425%. Model 4 has the best precision reaching 99.01%, but it has recall of 89.44%. PCXRNet achieves a recall of 96.838% and outperforms model 4 by 7.398%. It can be seen from the above analysis that PCXRNet has strong stability and achieves remarkable performance in recognition of pneumonia images, compared with other state-of-the-art CNNs.

H. Results on Other pneumonia Datasets

In this section, we verify the effectiveness of PCXRNet on other pneumonia datasets.

(1) Results on COVID-19 Radiography Dataset

Rahman et al. [43] first used the modified U-Net to segment lung area and then applied a classification network to recognize the COVID-19. This experiment did not use other data enhancement. Table 13 shows that PCXRNet achieves the best performance in accuracy, recall, precision and F1-score of 94.132%, 94.403%, 94.682% and 94.682%, respectively. Fig. 9(a) is the confusion matrix of PCXRNet on the COVID-19

dataset. Compared with the deeper ResNet50 and ResNet101, PCXRNet makes better use of the information on feature maps and strengthens the generalization ability. It can be seen from this experiment that PCXRNet not only displays a good classification performance on the childhood pneumonia dataset but also achieves remarkable performance on the COVID-19 dataset.

TABLE 12

THE COMPARISONS OF OTHER STATE-OF-THE-ART NETWORKS WITH OUR METHOD ON CHESTXRAY2017 DATASET

	model	Acc/%	Rec/%	Pre/%	F1/%
1	VGG16 [50]	86.26	85.22	87.73	86.46
2	VGG19 [50]	85.94	90.43	80.39	85.11
3	InceptionV3 [50]	94.59	95.35	93.75	94.54
4	DensNet201 [50]	93.66	89.44	99.01	93.98
5	Resnet50 [50]	96.61	94.92	98.49	96.67
6	Liang et al. [51]	90.50	96.70	89.10	92.70
7	Kermay et al. [22]	92.8	93.2	-	-
8	Hu et al. [23]	93.4	97.4	-	-
9	Chouhan et al. [52]	96.39	99.62	93.28	-
10	PCXRNet (ours)	97.436	96.838	97.705	97.241

TABLE 13

THE COMPARISONS OF OTHER STATE-OF-THE-ART NETWORKS WITH OUR METHOD ON COVID-19 DATASET

	model	Acc/%	Rec/%	Pre/%	F1/%
1	ResNet18 [43]	93.43	93.43	93.43	93.42
2	ResNet50 [43]	93.01	93.02	93.12	93.04
3	ResNet101 [43]	93.01	93.01	93.04	93.0
4	ChexNet [43]	93.21	93.21	93.28	93.20
5	DenseNet201 [43]	92.7	92.7	92.78	92.72
6	InceptionV3 [43]	93.46	93.47	93.49	93.47
7	PCXRNet (ours)	94.132	94.403	94.682	94.539

Fig. 9. Confusion matrix of other chest X-ray images datasets. The confusion matrix of COVID-19, and Tuberculosis datasets are on the left, and right, respective

(2) Results on Tuberculosis Chest X-ray Database

It is worth note that we do not use any image augmentation in this part. The input images were resized to 224×224 pixels. The initial learning rate was 0.001 and the maximum epoch was 100. The images were trained using the batch size of 16 with SGD

optimizer. Fig. 9(b) is the confusion matrix of PCXRNet on the Tuberculosis dataset. Table 14 shows that PCXRNet achieves state-of-the-art results on the Tuberculosis dataset, compared with traditional CNNs.

TABLE 14

THE COMPARISONS OF OTHER STATE-OF-THE-ART NETWORKS WITH OUR METHOD ON TUBERCULOSIS DATASET

	model	Acc/%	Rec/%	Pre/%	F1/%
1	VGG19 [44]	95.8	95.8	95.95	95.8
2	ResNet18 [44]	93.85	93.85	94.08	93.84
3	ResNet50 [44]	93.11	93.11	93.4	93.16
4	ResNet101 [44]	94.55	94.55	94.74	94.54
5	InceptionV3 [44]	95.72	95.73	95.92	95.73
6	DenseNet201 [44]	95.07	95.07	95.27	95.07
7	MobileNet [44]	94.33	94.33	94.65	94.32
8	ChexNe [44]	96.47	96.47	96.62	96.47
9	PCXRNet (ours)	99.357	99.357	99.357	99.357

V. CONCLUSION

In this paper, we proposed a novel deep learning network named PCXRNet, which contains two attention modules: CDSE and MCSA. Unlike traditional channel attention modules, CDSE first condenses the global information channel by channel using grouped depthwise convolution and then extracted channel-wise attention. MCSA uses three consecutive convolutions with the smaller kernel size to replace the original convolution in order to reduce parameters and introduce more nonlinearity. Numerous ablation experiments strongly proved that CDSE and MCSA are two complementary attention modules, where CDSE focuses on extracting channel-wise attention weights to exploit the inter-channel relationship of feature maps, while MCSA mainly extracts spatial-wise attention weights to capture disease-specific information from feature maps. For the qualitative analysis, we visualize the output of PCXRNet by using the Grad-CAM algorithm. Results show that our proposed method can utilize the lesion information in the chest X-ray images more effectively. We conduct extensive experiments on three chest x-ray recognition datasets to verify the effectiveness of our proposed method, which consistently achieves the state-of-the-art result.

REFERENCES

- [1] K. Watkins and S. Devi, "Pneumonia: a global cause without champions," *Lancet*, vol. 392, no. 10149, pp. 718-719, 2018.
- [2] R. Le, M. David, and J. Z. Heather, "Community-acquired pneumonia in children—a changing spectrum of disease," *Pediatric radiology*, vol. 47, no. 11, pp. 1392-1398, 2017.
- [3] I. J. Haq, A. C. Battersby, K. Eastham, and M. Michael, "Community acquired pneumonia in children," *Bmj*, vol. 356, 2017.
- [4] D. A. McAllister, et al., "Global, regional, and national estimates of pneumonia morbidity and mortality in children younger than 5 years between 2000 and 2015: a systematic analysis," *The Lancet Global Health*, vol. 7, no. 1, pp. e47-e57, 2019.
- [5] K. L. O'Brien et al., "Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the

- PERCH multi-country case-control study," *The Lancet*, vol. 394, no. 10200, pp. 757-779, 2019.
- [6] A. Abbas, M. A. Mohammed, and M. G. Mohamed, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *Applied Intelligence*, vol. 51, no. 2, pp. 854-864, 2021.
 - [7] E. A. Krupinski, K. S. Berbaum, R. T. Caldwell, K. M. Scharz, and J. Kim, "Long radiology workdays reduce detection and accommodation accuracy," *Journal of the American College of Radiology*, vol. 7, no. 9, pp. 698-704, 2010.
 - [8] M. Khalil, H. Ayad, and A. Adib, "MR-brain image classification system based on SWT-LBP and ensemble of SVMs," *International Journal of Medical Engineering and Informatics*, vol. 13, no. 2, pp. 129-142, 2021.
 - [9] M. M. Ghiasi, and S. Zendeheboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Computers in Biology and Medicine*, vol. 128, pp. 104089, 2021.
 - [10] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
 - [11] Y. Su, D. Li, and X. D. Chen, "Lung nodule detection based on faster R-CNN framework," *Computer Methods and Programs in Biomedicine*, vol. 200, pp. 105866, 2021.
 - [12] A. Saood and I. Hatem, "COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet," *BMC Medical Imaging*, vol. 21, no. 1, pp. 1-10, 2021.
 - [13] N. Lessmann, et al., "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 615-625, 2017.
 - [14] M. Aledhari, S. Jaji, M. Hefaida, and F. Saeed, "Optimized CNN-based Diagnosis System to Detect the Pneumonia from Chest Radiographs," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019.
 - [15] N. H. Jamal, I. S. Sayed, and W. S. Syed, "Estimation of organ absorbed dose in pediatric chest X-ray examination: A phantom study," *Radiation Physics and Chemistry*, vol. 166, pp. 108472, 2020.
 - [16] T. Kretz, K. R. Müller, T. Schaeffter, and C. Elster, "Mammography image quality assurance using deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 12, pp. 3317-3326, 2020.
 - [17] W. Qian, et al., "Identification of Melanoma From Hyperspectral Pathology Image Using 3D Convolutional Networks," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 218-227, 2020.
 - [18] I. Imran, M. Younus, K. Walayat, M. U. Kakar, and J. W. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Computerized Medical Imaging and Graphics*, vol. 88, pp. 101843, 2021.
 - [19] K. Abbas, H. Kim, and L. Chua, "PMED-Net: Pyramid Based Multi-Scale Encoder-Decoder Network for Medical Image Segmentation," *IEEE Access*, vol. 9, pp. 55988-55998, 2021.
 - [20] M. F. J. Acosta, L. Y. C. Tovar, M. B. G. Zapirain, and W. S. Percybrooks, "Melanoma diagnosis using deep learning techniques on dermatoscopic images," *BMC Medical Imaging*, vol. 21, no. 1, pp. 1-11, 2021.
 - [21] S. W. Chung, et al., "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm," *Acta orthopaedica*, vol. 89, no. 4, pp. 468-473, 2018.
 - [22] D. S. Kermany, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.
 - [23] M. J. Hu, et al., "Learning to recognize chest-Xray images faster and more efficiently based on multi-kernel depthwise convolution," *IEEE Access*, vol. 8, pp. 37265-37274, 2020.
 - [24] M. Masud, et al., "A Pneumonia Diagnosis Scheme Based on Hybrid Features Extracted from Chest Radiographs Using an Ensemble Learning Algorithm," *Journal of Healthcare Engineering*, 2021.
 - [25] B. Wang, et al., "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system," *Applied Soft Computing*, vol. 98, pp. 106897, 2020.
 - [26] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Computers in Biology and Medicine*, vol. 132, pp. 104348, 2021.
 - [27] L. d. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020.
 - [28] J. P. Zhang, Y. T. Xie, Y. Li, C. H. Shen, and Y. Xia, "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv:2003.12338*, 2020.
 - [29] A. MacNeil, P. Glaziou, C. Sismanidis, A. Date, S. Maloney, and K. Floyd, "Global epidemiology of tuberculosis and progress toward meeting global targets—worldwide, 2018," *Morbidity and Mortality Weekly Report*, vol. 69, no. 11, pp. 281-285, 2020.
 - [30] P. Lakhani, and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574-582, 2017.
 - [31] T. K. Kim, P. H. Yi, G. D. Hager, and C. T. Lin, "Refining dataset curation methods for deep learning-based automated tuberculosis screening," *Journal of Thoracic Disease*, vol. 1, no. 9, pp. 5078, 2020.
 - [32] T. Rahman, et al., "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8 pp. 191586-191601, 2020.
 - [33] L. Li, et al., "A large-scale database and a CNN model for attention-based glaucoma detection," *IEEE transactions on medical imaging*, vol. 39, no. 2, pp. 413-424, 2019.
 - [34] J. Fu, W. S. Li, J. Du, and Y. P. Huang, "A multiscale residual pyramid attention network for medical image fusion," *Biomedical Signal Processing and Control*, vol. 66, pp. 102488, 2021.
 - [35] R. Gu, et al., "CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699-711, 2020.
 - [36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018.
 - [37] Q. L. Zhang, and Y. B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2235-2239, 2021.
 - [38] B. C. Li, G. X. Kang, K. Cheng, and N. B. Zhang, "Attention-guided convolutional neural network for detecting pneumonia on chest x-rays," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4851-4854, 2019.
 - [39] O. Y. Xi, et al., "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2595-2605, 2020.
 - [40] S. W. Liang, Z. Z. Huang, M. F. Liang, and H. Z. Yang, "Instance enhancement batch normalization: An adaptive regulator of batch noise," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 4819-4827, 2020.
 - [41] D. Misra, "Mish: A self regularized non-monotonic neural activation function," *arXiv:1908.08681*, 2019.
 - [42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19, 2018.
 - [43] T. Rahman, et al., "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in biology and medicine*, vol. 132, pp. 104319, 2021.
 - [44] T. Rahman, et al., "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8 pp. 191586-191601, 2020.
 - [45] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Detrac: Transfer learning of class decomposed medical images in convolutional neural networks," *IEEE Access*, vol. 8, pp. 74901-74913, 2020.
 - [46] B. P. Health. Belarus Tuberculosis Portal. [Online]. Available: <http://tuberculosis.by>
 - [47] A. Rosenthal, et al., "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *Journal of clinical microbiology*, vol. 55, no. 11, pp. 3267-3282, 2017.
 - [48] Kaggle RSNA Pneumonia Detection Challenge [Online]. Available: <https://www.kaggle.com/c/rsna-pneumoniadetection-challenge/data>.
 - [49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618-626, 2017.
 - [50] K. E. Asnaoui, Y. Chawki, and A. Idri, "Automated methods for detection and classification pneumonia based on x-ray images using deep learning," *arXiv:2003.14363*, 2020.
 - [51] G. B. Liang, and L. X. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Computer methods and programs in biomedicine*, vol. 187, pp. 104964, 2020.
 - [52] V. Chouhan, et al., "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Applied Sciences*, vol. 10, no. 2, pp. 559, 2020.