Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01445-9

ORIGINAL ARTICLE

An attention mechanism module with spatial perception and channel information interaction

Yifan Wang[1] · Wu Wang[2,3] · Yang Li[1] · Yaodong Jia[1] · Yu Xu[1] · Yu Ling[1] · Jiaqi Ma[1]

Received: 29 November 2023 / Accepted: 31 March 2024

Abstract

In the field of deep learning, the attention mechanism, as a technology that mimics human perception and attention processes, has made remarkable achievements. The current methods combine a channel attention mechanism and a spatial attention mechanism in a parallel or cascaded manner to enhance the model representational competence, but they do not fully consider the interaction between spatial and channel information. This paper proposes a method in which a space embedded channel module and a channel embedded space module are cascaded to enhance the model's representational competence. First, in the space embedded channel module, to enhance the representational competence of the region of interest in different spatial dimensions, the input tensor is split into horizontal and vertical branches according to spatial dimensions to alleviate the loss of position information when performing 2D pooling. To smoothly process the features and highlight the local features, four branches are obtained through global maximum and average pooling, and the features are aggregated by different pooling methods to obtain two feature tensors with different pooling methods. To enable the output horizontal and vertical feature tensors to focus on different pooling features simultaneously, the two feature tensors are segmented and dimensionally transposed according to spatial dimensions, and the features are later aggregated along the spatial direction. Then, in the channel embedded space module, for the problem of no cross-channel connection between groups in grouped convolution and for which the parameters are large, this paper uses adaptive grouped banded matrices. Based on the banded matrices utilizing the mapping relationship that exists between the number of channels and the size of the convolution kernels, the convolution kernel size is adaptively computed to achieve adaptive cross-channel interaction, enhancing the correlation between the channel dimensions while ensuring that the spatial dimensions remain unchanged. Finally, the output horizontal and vertical weights are used as attention weights. In the experiment, the attention mechanism module proposed in this paper is embedded into the MobileNetV2 and ResNet networks at different depths, and extensive experiments are conducted on the CIFAR-10, CIFAR-100 and STL-10 datasets. The results show that the method in this paper captures and utilizes the features of the input data more effectively than the other methods, significantly improving the classification accuracy. Despite the introduction of an additional computational burden (0.5 M), however, the overall performance of the model still achieves the best results when the computational overhead is comprehensively considered.

Keywords Space embedded channel module · Horizontal weights · Vertical weights · Channel embedded space module · Banded matrices

Introduction

The attention mechanism allocates different weights to different input elements, enabling the model to focus more accurately on important information, thereby improving the model performance. In practical applications, attention mechanisms are widely used in fields such as natural language processing [1], computer vision [2], and image processing [3].

✉ Yang Li
lyang@cust.edu.cn

1 School of Electric and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

2 Northern Navigation Control Technology Co., Ltd, 2 Kechuang 15th Street, Yizhuang Economic and Technological Development Zone, Beijing 100000, China

3 School of Computing, Changchun University of Science and Technology, Changchun 130022, China

Published online: 06 May 2024

The channel attention mechanism [4–7] can adjust the feature weight based on the importance of different channels, making the model more focused on features that are beneficial to the task. The first proposed channel attention mechanism module, SE-Net [4], performs channel attention on the channel dimension relative to global operations. Based on the SE-Net, ECA-Net [5] and GCT [6] add interrelationships between channels. FCA-Net [7] extends the global average pooling operation in the channel attention mechanism to the two-dimensional discrete cosine transform form, increasing feature diversity.

The computational cost of the channel attention mechanism is relatively low, but the importance of spatial location is ignored, resulting in insufficient attention given to local features.

The spatial attention mechanism [8–10] assigns different weights to information from different positions, and the model focuses more on regions that are more important to the task. RAM [8] builds a spatial attention mechanism based on recurrent neural networks to enable the model to focus on the most relevant parts when processing the input sequence. STN [9] improves the accuracy of image classification via adaptive spatial transformation. CCNet [10] enhances the semantic information of features by establishing remote dependency relationships based on spatial attention. The spatial attention mechanism ignores the channel information of the input sequence or image, focusing too much on local information and ignoring the global context. Additionally, when processing natural language text with unstructured data, it is not as effective as when processing structured data such as images.

The mixed attention mechanism [11–17] combines channel and spatial information to capture multiscale features and improve model representational competence. BAM [11] and HAM [12] cascade channels and spatial attention mechanisms. The parallel channel and spatial attention mechanism of CBAM [13] and scSE [14] enhance the feature expression ability of convolutional neural networks. SA-Net [15] and EPSA-Net [16] use convolutional kernels of different sizes to process input feature maps before parallel connection, allowing the model to have multiscale feature information. Coordinated attention (CA) [17] cleverly embeds spatial dimension information into the channel dimension and aggregates features in two spatial directions while preserving accurate positional information and spatial information from different directions.

In summary, the channel attention mechanism [4–7] and the spatial attention mechanism [8–10] focus on information from different dimensions, respectively, which can affect the attention given to local features. The mixed attention mechanism [11–17] enhances the representational competence of the model through parallel or cascading; however, the introduction of mixed attention increases model complexity and requires carefully choosing how channels and space are combined.

To solve the above problems, this paper proposes a novel and effective attention mechanism module, namely, the space and channel mutually embedded attention mechanism. By designing a cascade with a space embedded channel module and a channel embedded space module, this paper comprehensively considers horizontal and vertical information in the channel dimension. By introducing an adaptive channel interaction mode, the channel correlation in different spatial dimensions is strengthened, thus solving the problem of channel and spatial attention mechanisms focusing on information from different dimensions. The design of mutual embedding not only enhances the model's comprehensive representation of local features, but also avoids the complexity of introducing mixed attention and eliminates the need to finely select how channels and space are combined.

In the space embedded channel module, first, the input intermediate feature tensor is split into two feature tensors in the horizontal ($C \times 1 \times W$) and vertical directions ($C \times H \times 1$), which alleviates the loss of position information when performing 2D pooling. Next, to introduce the features extracted by different pooling operations to increase the model's diversity for the input data, in this paper, the two feature maps are sequentially through 2D maximum and average pooling to obtain four sets of feature tensors (($C \times 1 \times W)_{avg}$, $(C \times 1 \times W)_{max}$, $(C \times H \times 1)_{avg}$ and $(C \times H \times 1)_{max}$). Then, to fuse the features extracted in the horizontal and vertical directions to capture the information of the input data more comprehensively, this paper merges the features in the horizontal and vertical directions according to different pooling methods to obtain two sets of feature tensors (($C \times 1 \times (W + H))_{avg}$ and $(C \times 1 \times (W + H))_{max}$). Finally, to preserve the different spatial dimension information in the location, the feature tensors (($C \times 1 \times (W + H))_{avg}$ and $(C \times 1 \times (W + H))_{max}$) are split and subsequently merged in different directions to obtain two sets of feature tensors. Moreover, each set of feature tensors is enabled to focus on the important regions of the module while enhancing the semantic information of the spatial location features.

In the channel embedded space module, first, the feature vectors of the horizontal and vertical dimensions output from the space embedded channel module are considered as two parallel inputs, and tensor transformation is performed to convert the two-dimensional tensor into a one-dimensional tensor. Next, to strengthen the correlation between channel dimensions in different spatial dimensions, banded matrices are introduced on the basis of grouped convolution to reduce the number of parameters, and the convolution kernel setting is approximated to be consistent with the input and output dimensions of the ECA-Net method. Subsequently, the two sets of feature tensors output from the channel adaptive interaction module of the 1D banded matrix are extended to the

spatial dimension by unsqueezing to obtain the channel features. Afterwards, a Sigmoid operation is performed to adjust the attention weights, limiting their range to between (0,1) to generate the attention weights of the channel dimensions in different spatial dimensions. Then, the expansion operation is used to expand in the height and width dimensions to match the size of the input original feature tensor in preparation for subsequent elementwise multiplication operations. Finally, an elementwise multiplication operation is used to output the result of the input original tensor adjusted with attention weights in the horizontal and vertical directions, allowing the model to more adaptively focus on information from different positions and spatial dimensions in the input tensor.

Main contributions:

(1) To our knowledge, this paper is the first to interactively embed channel and spatial attention. The interaction between channel and spatial attention enhances the model's representational competence.
(2) A space embedded channel module is constructed to enhance the representational competence for objects of interest. This module embeds the horizontal and vertical directions into the channel dimension, smooths image information and highlights local features through global maximum and average pooling, thus comprehensively considering feature information from different directions.
(3) A channel embedded space module is constructed, using an adaptive grouped banded matrix to enhance the correlation between channels in different spatial dimensions. The attention weights of the generated channel dimensions in different spatial dimensions are utilized to multiply elementwise with the original feature tensor to adjust the input tensor and make the model more adaptive in focusing on information from different channels and spatial dimensions.

Related work

In this section, a brief overview of the image classification network architecture based on convolutional neural networks is provided. A detailed review of the algorithm inspiration source in this paper, the CA [17] attention mechanism. The algorithm in this paper is proposed based on the analysis of the algorithm shortcomings.

Network engineering

"Network engineering" plays an important role in visual research, and algorithms based on convolutional neural networks such as LeNet [18], AlexNet [19], VGG [20], Inception [21], ResNet [22] and MobileNet [23–25] are commonly used in tasks such as image classification, object detection, and image segmentation in visual research. The LeNet [18] algorithm demonstrates the effectiveness of convolutional neural networks in image classification tasks. AlexNet [19] introduces a deeper network architecture and adopts ReLU activation functions and Dropout regularization to avoid gradient vanishing and overfitting problems. Its design and innovation lay the foundation for more complex networks such as VGG [20] and ResNet [22]. The VGG [20] network structure is relatively simple, with a 16–19 layer deep model. ResNet [22] solves the gradient vanishing and exploding problems in deep network training and can still achieve better performance even when the network depth exceeds 100 layers. Typical examples include ResNet18, ResNet56, ResNet110 and ResNet152. The MobileNet [23–25] series of algorithms is suitable for efficient and lightweight neural networks in mobile devices and embedded systems to achieve better performance when computing resources are limited. MobileNetV1 [23] replaces the convolutions in VGG [20] with deep separable convolutions, using ReLU6 as the activation function. MobileNetV2 [24] adds shortcut connections and expands the dimensionality, and the convolution uses linear activation instead of ReLU on the output pointwise. MobileNetV3 [25] introduces the inverted residual module and squeeze-and-excitation module based on V2 [24], using Hard-swish as the activation function.

In the experimental section of this paper, MobileNetV2 and various depths of ResNet are chosen to validate the attention module, evaluate its applicability to lightweight networks (such as mobile devices) and more powerful networks (such as servers), and determine its effectiveness in different environments.

Review and problem analysis of CA

The channel attention mechanism [4–7, 26] is an expression of feature abstraction, while the spatial attention mechanism [8–10, 27] is an enrichment of positional information. A single attention mechanism cannot simultaneously satisfy the acquisition of channel and positional information. Therefore, researchers have proposed a mixed attention mechanism [28], which generates multiple attention feature maps from multiple attention mechanisms and then concatenates them [11, 12, 29] or parallels them [13, 14, 30] to obtain richer feature representations. CA [17] cleverly attaches spatial information to channels, which can be plug-and-played on lightweight classification networks [31] with negligible computational overhead. CA [17] provides a new approach for mixed attention mechanisms.

The CA [17] algorithm first converts a three-dimensional tensor $F = [F_1, F_2,\ldots,F_C] \in R^{H\times W\times C}$ into two two-dimensional tensors $Z_C^H \in \mathbb{R}^{C\times H\times 1}$ and $Z_C^W \in \mathbb{R}^{C\times 1\times W}$ and computes the correlation between different dimensions in Eqs. (1) and (2).

$$Z_C^H(H) = \frac{1}{W}\sum_{0\le i\le W} X_C(H, i),\ Z_C^H \in \mathbb{R}^{C\times H\times 1} \tag{1}$$

where C denotes the number of channels, H denotes the height, and 1 denotes the width. Each channel corresponds to a feature at a different vertical position. In height, each channel represents different spatial information at different horizontal positions.

$$Z_C^W(W) = \frac{1}{H}\sum_{0\le j\le H} X_C(j, W),\ Z_C^W \in \mathbb{R}^{C\times 1\times W} \tag{2}$$

where 1 denotes the height and W denotes the width. Each channel corresponds to a feature at a different horizontal position. At the width position, each channel represents different spatial information at different horizontal positions.

The results are concatenated after global average pooling of $Z_C^H \in \mathbb{R}^{C\times H\times 1}$ and $Z_C^W \in \mathbb{R}^{C\times 1\times W}$, then split backward to obtain the new $Z_C^H \in \mathbb{R}^{C\times H\times 1}$ and $Z_C^W \in \mathbb{R}^{C\times 1\times W}$ after performing dimensionality reduction on the channels. Two feature maps are obtained through Eq. (3), and the output of the attention mechanism is the product of the feature maps.

$$\begin{cases} g_H = Sigmoid(F_H(\mathrm{Re}LU(F_{1*1}(Z_C^H, Z_C^W)) \\ g_W = Sigmoid(F_W(\mathrm{Re}LU(F_{1*1}(Z_C^H, Z_C^W)) \end{cases} \tag{3}$$

In the CA algorithm, the idea of splitting spatial dimension information into two dimensions and embedding them into channels is very novel and effective. However, the following issues remain:

(1) Only global average pooling is performed after spatial dimension splitting, and the loss of detailed local feature information in the feature map makes it more difficult for the network to capture the local structure in the image.
(2) In the paper, only spatial dimension information is embedded into the channels. Can channel information be embedded in the spatial dimension?
(3) Processing channel information by dimensionality reduction leads to data loss, while focusing primarily on local interchannel relationships fails to capture longer-range dependencies.

In response to the above issues, this paper proposes an attention mechanism module with channel and spatial dimension information interaction. A combination of multiple pooling units is used to provide greater feature richness than single pooling; channels are divided into multiple groups using banded matrices, treating each group as a specific type of channel information. Keeping the spatial dimension unchanged, each pixel can capture channel information from different groups, thus embedding channel information into the spatial dimension and avoiding the data loss problem that would be caused by processing channel information in a reduced-dimensional way through banded matrices. In promoting information exchange between different dimensions, more correlations are introduced into the model to enhance feature representation competence.

Methodology

Assuming $F = [F_1, F_2,\ldots, F_C] \in R^{C\times H\times W}$ is used as the intermediate feature map of the input tensor, the output is $F' = [F_1', F_2',\ldots, F_C'] \in R^{C\times H\times W}$. A schematic diagram of the attention mechanism proposed in this paper is shown in Fig. 1. The SPCII achieves interaction between channel and spatial dimension information through two steps: embedding spatial dimension information into the channel dimension and embedding channel dimension information into the spatial dimension. The following provides a detailed description of SPCII.

Space embedded channel module

As shown in Fig. 1, the red solid line on the left side shows the spatial dimension information embedded in the channel module. Referring to the CA [17] algorithm, Z_C^H and Z_C^W are obtained, and “where” is embedded into "what". That is, the channel dimension remains unchanged and still represents different channel features; the spatial dimensions of the horizontal and vertical directions are compressed to 1, and different position information is embedded into the channel dimension.

Because only average pooling is used in reference [17] to preserve smooth image information, this paper refers to the CBAM [13] attention mechanism to add maximum pooling on the basis of average pooling, such as Eqs. (4) and (5). The combination of the two helps to highlight local features while preserving spatial smooth information.

$$\begin{cases} F_{\mathrm{Avg}}^H = AvgPool(Z_C^H(H)) \\ F_{Max}^H = MaxPool(Z_C^H(H)) \end{cases} \tag{4}$$

$$\begin{cases} F_{Avg}^W = AvgPool(Z_C^W(W)) \\ F_{Max}^W = MaxPool(Z_C^W(W)) \end{cases} \tag{5}$$

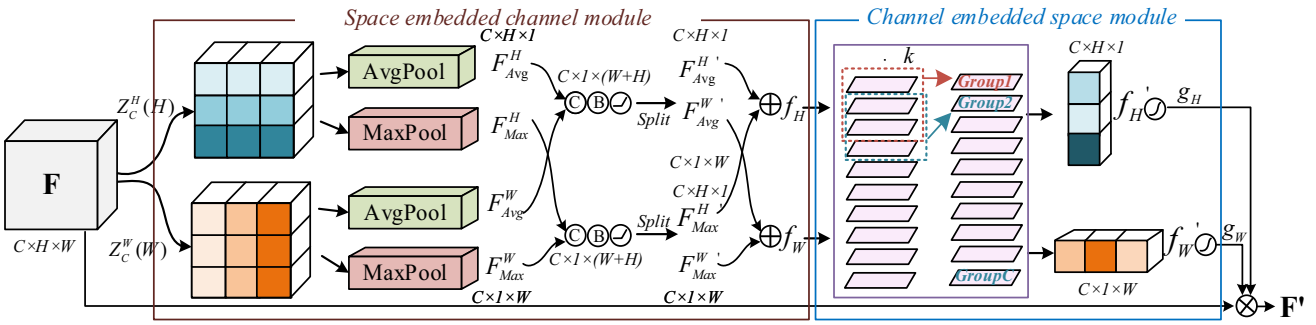

Fig. 1 Schematic diagram of the SPCII module

Then, two feature tensors obtained by the same pooling method are combined in the horizontal and vertical directions to form two new feature tensors $Concat(F^{W}_{Avg}, F^{H}_{Avg}) \in \mathbb{R}^{C\times1\times(H+W)}$ and $Concat(F^{W}_{Max}, F^{H}_{Max}) \in \mathbb{R}^{C\times1\times(H+W)}$. The model simultaneously considers feature information from different directions.

After Eq. (6) is applied, $F^{HW'}_{Avg}$ and $F^{HW'}_{Max}$ are obtained through 2D convolution, batch normalization and activation function processing to obtain a richer output with higher-level feature representation.

$$\begin{cases} F^{HW'}_{Avg} = \mathrm{ReLU}(BN(Conv2D(Concat(F^{W}_{Avg}, F^{H}_{Avg})))) \\ F^{HW'}_{Max} = \mathrm{ReLU}(BN(Conv2D(Concat(F^{H}_{Max}, F^{H}_{Max})))) \end{cases} \tag{6}$$

Finally, $F^{HW'}_{Avg}$ and $F^{HW'}_{Max}$ are segmented to obtain $F^{H'}_{\mathrm{Avg}}$, $F^{W'}_{Avg}$, $F^{H'}_{Max}$ and $F^{W'}_{Max}$. Note that the dimensions of $F^{W'}_{Avg}$ and $F^{W'}_{Max}$ are transposed, and the horizontal and vertical dimensions are swapped. The channels are then summed to improve the ability of each branch to capture image features in Eq. (7).

$$\begin{cases} f_H = F^{H'}_{\mathrm{Avg}} + F^{H'}_{Max} \\ f_W = F^{W'}_{Avg} + F^{W'}_{Max} \end{cases} \tag{7}$$

In Eq. (7), the output of f_H is $(C, (H_A + H_M), 1)$ and the output of f_W is $(C, 1, (W_A + W_M))$.

Channel embedded space module

As shown in Fig. 1, the blue solid line on the right side shows the channel embedded space module. This module takes the feature vectors in the horizontal and vertical dimensions output by the space embedded channel module as two parallel inputs.

The inputs f_H and f_W are both two-dimensional tensors. To embed the channels into different spatial dimensions, taking f_H as an example, this paper defines a new tensor, $new\_h = (H_A + H_M) \times 1$, that shares data storage with the original tensor, where each channel contains elements at the corresponding positions in the original f_H. The same goes for f_W.

A tensor view transformation operation using Eq. (8) adapts the input requirements of a one-dimensional convolution (C, new_h) with a convolution kernel of k, ensuring that the input shape is correct.

$$\begin{cases} f_H.view(f_H.size(0), new\_h) \\ f_W.view(f_W.size(0), new\_w) \end{cases} \tag{8}$$

The f_H.view(…) tensor view transformation operation in PyTorch is used and the view function acts to change the tensor shape without changing the tensor elements. In Eq. (8), f_H.size(0) denotes the channel dimension and $new\_h = (H_A + H_M) \times 1$ denotes the height dimension.

This paper obtains the number of channels by $C = f_H$.size(0), which is used as the number of input and output channels for 1D convolution.

With respect to the convolution kernel k, the 1D banded matrix for adaptive cross-channel interactions (ACCI) from the ECA-Net [4] algorithm is used to strengthen the correlation between channel dimensions. The 1D convolution is changed into a grouped convolution, which is divided into several blocks, and each block is fully connected internally with a parameter C^2. There is no cross-channel connection between groups, and the parameter is larger (the parameter is C^2/G), where G represents the number of groups in the grouped convolution. Therefore, the banded matrix is introduced as in Eq. (9), which reduces the parameter size (the parameter at this point is $k \times C$) while keeping the input and output dimensions consistent and enhancing the correlation between channels.

$$w_G = \begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{9}$$

There exists a mapping between C and k in Eq. (10), and the relationship $C = 2^{r\times k-b}$ is defined in reference [4]. To ensure that the convolution kernel has a center point, the convolution kernel size is forced to be an odd number and the convolution kernel size is adaptively computed as k by Eq. (10).

$$\begin{cases} k\%2 = 0 \Rightarrow k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right| - 1 \\ k\%2 = 1 \Rightarrow k = \psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right| \end{cases} \tag{10}$$

where $r = 2$ and $b = 1$.

Taking (C,k,1) as an input to the ACCI outputs a new one-dimensional tensor C, namely, the relationship between the channels.

In this paper, to extend the results of the ACCI to the spatial dimension to achieve a channel embedded spatial dimension, we use.unsqueeze(-1) to add a dimension to the last dimension, i.e., (C, (H_A + H_M), 1), matching the original two-dimensional tensor. In the f_W part, we use.unsqueeze(–2) to add a dimension in the second to last position, i.e., (C, 1, (W_A + W_M)). The attention weights are adjusted by a Sigmoid operation (the Sigmoid activation function in Eq. (11)), which adjusts the range of the attention weights and limits them to between (0, 1). The attention weights of the channel dimension in different spatial dimensions are generated. The correlation between different channels is dynamically adjusted by the weights of the banded matrix, allowing the model to focus more flexibly on important channel information in different spatial dimensions.

$$\begin{cases} g_H = Sigmoid(Conv1D\_k(f_H)) \\ g_W = Sigmoid(Conv1D\_k(f_W)) \end{cases} \tag{11}$$

The outputs g_H and g_W are the channel information about which positions the model should focus on horizontally and vertically, respectively; that is, different channel information is embedded in the spatial dimension. The.expand operation is used to expand g_H and g_W in the height and width dimensions to match the dimension of the input F_C for subsequent elementwise multiplication operations. The output of the attention block $F'_C(i,\ j)$ is given by Eq. (12).

$$F'_C(i,\ j) = F_C(i,\ j) \times g_C^H(i) \times g_C^W(j) \tag{12}$$

The output of the attention block is the result of the input tensor F_C being adjusted by the attention weights g_H and g_W in the horizontal and vertical directions, respectively. The attention block can more adaptively focus on information from different positions and spatial dimensions in the input tensor to improve performance in image processing tasks.

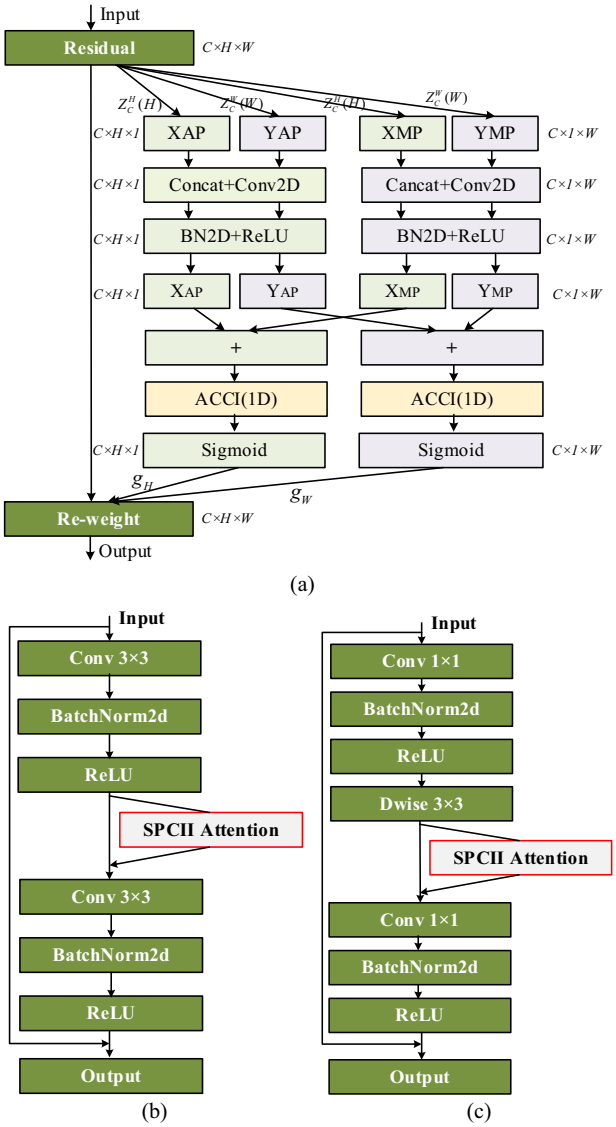

Fig. 2 SPCII Attention Generation. **a** SPCII attention **b** MobileNetV2 + SPCII attention **c** ResNet + SPCII attention

SPCII attention generation

Figure 2a shows the SPCII attention mechanism module proposed in this paper, which can be plug-and-played in any CNN architecture, where XAP and YAP are the average pooling operations in the horizontal and vertical directions, XMP and YMP are the maximum pooling operations in the horizontal and vertical directions, and X_{AP}, Y_{AP}, X_{MP} and Y_{MP} are the tensors in the horizontal and vertical directions after splitting. The ACCI is the module for adaptive cross-channel interactions in Sect. "Channel embedded space module". Figure 2b shows the integration schematic of SPCII and MobileNetV2 [24], and Fig. 2c shows the integration schematic of SPCII and ResNet [22] (BasicBlock for example).

Experiments

Experiment setup

This paper implements experiments using the PyTorch toolkit on the VScode platform. To evaluate the algorithms of this paper, experiments are performed on standard datasets CIFAR-10, CIFAR-100, and STL-10. During training, a standard SGD optimizer is used with a decay rate of 0.9, a decay weight of 4E-5, and an initial learning rate of 0.05. MobileNetV2 is used as the baseline with 200 epochs and a batchsize of 64.

The CIFAR-10 [32] and CIFAR-100 [32] datasets both contain 60,000 RGB images at 32 × 32 resolution. Among them, 83.33% are used as the training set, and 16.67% are used as the test set. The CIFAR-10 dataset contains 10 categories, and the CIFAR-100 dataset contains 100 categories.

The STL-10 [33] dataset contains 113,000 RGB images with 96 × 96 resolution from ImageNet, and includes 10 categories. The training set contains 5000 images, the test set contains 8000 images, and the unlabeled set contains 100,000 unlabeled images. Only the training and test sets of STL-10 are used in the experiments of this paper, and the unlabeled dataset is not used.

To evaluate the performance advantages and disadvantages of the method proposed in this paper, five methods are compared: SE-Net (channel attention mechanism) [4], ECA-Net (lightweight channel attention mechanism) [5], CBAM (channel + spatial attention mechanism) [13], CA (space embedded channel attention mechanism) [17] and SPCII (the mechanism proposed in this paper). In addition, since no precedent using the same public dataset for validation has been found in previous research, this paper achieves consistency in the hardware environment and the network location of module insertion, and retrains the model. The Parameters, GFLOPs, and error rates reported in Tables 1, 2, 3, 4, 5, 6 are the average results of 10 runs in the same environment.

Image classification on the CIFAR datasets

We conduct target classification experiments on the CIFAR-10, STL-10 and CIFAR-100 datasets to evaluate the SPCII attention mechanism module, following the training rules and parameters mentioned in Sect. "Experiment Setup" and embed the SPCII module into the MobileNetV2 (lightweight network) and ResNetX (deep network) series of target classification networks. First, SPCII is embedded into the MobileNetV2 and ResNet18 backbone models. Then, the ResNet depth is increased to observe the robustness of the SPCII.

Table 3 Comparison between different CNN architectures on the STL-10 dataset

Description	Parameters (M)	GFLOPs	STL-10 Error (%)
MobileNetV2 (Baseline)	2.237	0.326284	34.76
MobileNetV2 + SE	**2.258$_{+0.94\%}$**	0.328$_{+0.53\%}$	34.96$_{+0.58\%}$
MobileNetV2 + ECA	2.273$_{+1.61\%}$	**0.352$_{+7.88\%}$**	40.52$_{+16.57\%}$
MobileNetV2 + CBAM	2.811$_{+25.66\%}$	0.345825$_{+5.99\%}$	38.45$_{+10.62\%}$
MobileNetV2 + CA	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	34.22$_{-1.55\%}$
MobileNetV2 + Ours	2.682$_{+19.89\%}$	0.339166$_{+3.95\%}$	**33.40$_{-3.91\%}$**
ResNet18 (Baseline)	12.577	1.826	39.34
ResNet18 + SE	11.187$_{-11.05\%}$	1.824$_{-0.11\%}$	36.38$_{-7.52\%}$
ResNet18 + ECA	**11.884$_{-5.51\%}$**	1.825$_{-0.05\%}$	38.36$_{-2.49\%}$
ResNet18 + CBAM	11.188$_{-11.04\%}$	1.825$_{-0.05\%}$	35.50$_{-9.76\%}$
ResNet18 + CA	11.256$_{-10.50\%}$	1.826$_{+0.00\%}$	35.88$_{-8.80\%}$
ResNet18 + Ours	11.256$_{-10.50\%}$	**1.827$_{+0.05\%}$**	**35.12$_{-10.73\%}$**

The best values under different descriptions are shown in bold

Table 5 Ablation experiment data for the space embedded channel module

Description	Parameter (M)	GFLOPs	CIFAR-10 Error (%)
MobileNetV2 (Baseline)	**2.237**	0.326284	17.98
+ CA	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	17.80$_{-1.00\%}$
+ Ours(M + C)	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	17.77$_{-1.17\%}$
+ Ours(A + C)	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	17.78$_{-1.11\%}$
+ Ours(M + A + C)	2.682$_{+19.89\%}$	**0.339166$_{+3.95\%}$**	**17.20$_{-4.33\%}$**

The best values under different descriptions are shown in bold

Comparison between different backbone models

The SPCII proposed in this paper is embedded into MobileNetV2 and ResNet18, and its performance is compared with that of representative SE (channel attention mechanism) [4], ECA (lightweight channel attention mechanism) [5], CBAM (channel + spatial attention mechanism)

Table 1 Comparison between different CNN architectures on the CIFAR-10 and CIFAR-100 datasets

Description	Parameters (M)	GFLOPs	CIFAR-10 Error (%)	CIFAR-100 Error (%)
MobileNetV2(Baseline)	2.237	0.326284	17.98	48.19
MobileNetV2 + SE	**2.258$_{+0.94\%}$**	0.328205$_{+0.59\%}$	17.53$_{-2.50\%}$	48.45$_{+0.54\%}$
MobileNetV2 + ECA	2.273$_{+1.609\%}$	**0.352.068$_{+7.90\%}$**	17.27$_{-3.95\%}$	48.02$_{-0.35\%}$
MobileNetV2 + CBAM	2.811$_{+25.66\%}$	0.345825$_{+5.99\%}$	17.60$_{-2.11\%}$	47.65$_{-1.12\%}$
MobileNetV2 + CA	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	17.80$_{-1.00\%}$	47.24$_{-1.97\%}$
MobileNetV2 + Ours	2.682$_{+19.89\%}$	0.339166$_{+3.95\%}$	**17.20$_{-4.33\%}$**	**47.07$_{-2.32\%}$**
ResNet18	11.182	1.824	14.77	44.61
ResNet18 + SE	11.277$_{+0.85\%}$	1.824$_{+0.00\%}$	14.09$_{-4.60\%}$	43.02$_{-3.56\%}$
ResNet18 + ECA	12.577$_{+12.48\%}$	1.826$_{+0.11\%}$	15.27$_{+3.39\%}$	43.30$_{-2.94\%}$
ResNet18 + CBAM	**11.256$_{+0.66\%}$**	1.826$_{+0.11\%}$	13.93$_{-5.69\%}$	42.29$_{-5.20\%}$
ResNet18 + CA	**11.256$_{+0.66\%}$**	1.826$_{+0.11\%}$	13.93$_{-5.69\%}$	42.22$_{-5.36\%}$
ResNet18 + Ours	**11.256$_{+0.66\%}$**	**1.827$_{+0.16\%}$**	**13.46$_{-8.87\%}$**	**42.06$_{-5.72\%}$**

The best values under different descriptions are shown in bold

Table 2 Comparison between ResNet architectures with different depths on the CIFAR-10 and CIFAR-100 datasets

Description	Parameters	GFLOPs	CIFAR-10 Error (%)	CIFAR-100 Error (%)
ResNet20	11.228	1.824	16.43	41.61
ResNet20 + SE	**11.234$_{+0.05\%}$**	1.824$_{+0.00\%}$	16.22$_{-1.28\%}$	41.02$_{-1.42\%}$
ResNet20 + ECA	12.623$_{+12.42\%}$	1.826$_{+0.11\%}$	16.21$_{-1.34\%}$	44.30$_{+6.46\%}$
ResNet20 + CBAM	11.318$_{+0.80\%}$	1.825$_{+0.05\%}$	16.11$_{-1.95\%}$	41.29$_{-0.77\%}$
ResNet20 + CA	11.303$_{+0.67\%}$	1.826$_{+0.11\%}$	16.08−$_{2.13\%}$	40.22$_{-3.34\%}$
ResNet20 + Ours	11.303$_{+0.67\%}$	**1.827$_{+0.16\%}$**	**15.99$_{-2.68\%}$**	**40.03$_{-3.80\%}$**
ResNet32	21.336	3.678	15.32	40.67
ResNet32 + SE	21.497$_{+0.75\%}$	3.680$_{+0.05\%}$	14.35$_{-6.33\%}$	38.28$_{-5.88\%}$
ResNet32 + ECA	23.857$_{+11.82\%}$	3.682$_{+0.11\%}$	15.22$_{-0.65\%}$	39.58$_{-2.68\%}$
ResNet32 + CBAM	**21.449$_{+0.53\%}$**	3.681$_{+0.08\%}$	14.27$_{-6.85\%}$	37.60$_{-7.55\%}$
ResNet32 + CA	21.471$_{+0.63\%}$	3.683$_{+0.14\%}$	**14.15$_{-7.64\%}$**	**36.04$_{-11.38\%}$**
ResNet32 + Ours	21.471$_{+0.63\%}$	3.684$_{+0.16\%}$	14.17$_{-7.51\%}$	36.47$_{-10.33\%}$
ResNet56	23.713	4.132	16.29	37.58
ResNet56 + SE	26.244$_{+10.67\%}$	4.140$_{+0.19\%}$	16.27$_{-0.12\%}$	37.39$_{-0.51\%}$
ResNet56 + ECA	**23.121$_{-2.50\%}$**	4.141$_{+0.22\%}$	16.33$_{+0.25\%}$	37.33$_{-0.67\%}$
ResNet56 + CBAM	23.571$_{-0.60\%}$	4.140$_{+0.19\%}$	15.27$_{-6.26\%}$	36.98$_{-1.60\%}$
ResNet56 + CA	23.793$_{+0.34\%}$	**4.160$_{+0.68\%}$**	15.32$_{-5.95\%}$	36.55$_{-2.74\%}$
ResNet56 + Ours	23.793$_{+0.34\%}$	**4.160$_{+0.68\%}$**	**15.22$_{-6.57\%}$**	**36.42$_{-3.09\%}$**
ResNet110	41.361	7.616	13.12	36.37
ResNet110 + SE	41.666$_{+0.74\%}$	7.618$_{+0.03\%}$	12.96$_{-1.22\%}$	36.22$_{-0.41\%}$
ResNet110 + ECA	41.613$_{+0.61\%}$	7.623$_{+0.09\%}$	12.4$_{-5.49\%}$	36.22$_{-0.41\%}$
ResNet110 + CBAM	**41.389$_{+0.07\%}$**	7.620$_{+0.05\%}$	12.77$_{-2.67\%}$	36.32$_{-0.14\%}$
ResNet110 + CA	41.610$_{+0.60\%}$	7.625$_{+0.12\%}$	12.84$_{-2.13\%}$	36.21$_{-0.44\%}$
ResNet110 + Ours	41.610$_{+0.60\%}$	**7.627$_{+0.14\%}$**	**12.00$_{-8.54\%}$**	**36.14$_{-0.63\%}$**

The best values under different descriptions are shown in bold

Table 4 Comparison between ResNet architectures with different depths on the STL-10 dataset

Description	Parameters	GFLOPs	STL-10 Error (%)
ResNet20	17.2506	1.319	35.33
ResNet20 + SE	**17.4088$_{+0.92\%}$**	1.322$_{+0.23\%}$	37.73$_{+6.79\%}$
ResNet20 + ECA	19.4346$_{+12.66\%}$	1.322$_{+0.23\%}$	35.83$_{+1.42\%}$
ResNet20 + CBAM	17.4682$_{+1.26\%}$	**1.335$_{+1.21\%}$**	34.17$_{-3.28\%}$
ResNet20 + CA	17.8474$_{+3.46\%}$	1.326$_{+0.53\%}$	34.88$_{-1.27\%}$
ResNet20 + Ours	17.8474$_{+3.46\%}$	1.326$_{+0.53\%}$	**33.62$_{-4.84\%}$**
ResNet34	21.116	7.1199	40.49
ResNet34 + SE	21.277$_{+0.76\%}$	7.1211$_{+0.02\%}$	36.95$_{-8.74\%}$
ResNet34 + ECA	**21.13$_{+0.07\%}$**	**7.1261$_{+0.09\%}$**	37.26$_{-7.98\%}$
ResNet34 + CBAM	21.27$_{+0.73\%}$	7.1244$_{+0.06\%}$	37.10$_{-8.37\%}$
ResNet34 + CA	21.251$_{+0.64\%}$	7.1233$_{+0.05\%}$	36.51$_{-9.83\%}$
ResNet34 + Ours	21.251$_{+0.64\%}$	7.1238$_{+0.05\%}$	**35.54$_{-12.23\%}$**
ResNet56	23.592	4.132	50.92
ResNet56 + SE	26.060$_{+10.46\%}$	4.140$_{+0.19\%}$	50.25$_{-1.32\%}$
ResNet56 + ECA	63.790$_{+170.39\%}$	4.177$_{+1.09\%}$	49.69$_{-2.42\%}$
ResNet56 + CBAM	26.061$_{+10.47\%}$	4.141$_{+0.22\%}$	49.80$_{-2.20\%}$
ResNet56 + CA	**25.446$_{+7.86\%}$**	4.170$_{+0.92\%}$	49.22$_{-3.34\%}$
ResNet56 + Ours	**25.446$_{+7.86\%}$**	**4.182$_{+1.21\%}$**	**49.02$_{-3.73\%}$**
ResNet101	41.361	7.616	41.00
ResNet101 + SE	**41.386$_{+0.06\%}$**	7.618$_{+0.03\%}$	56.00$_{+36.59\%}$
ResNet101 + ECA	43.757$_{+5.79\%}$	7.620$_{+0.05\%}$	47.623$_{+16.15\%}$
ResNet101 + CBAM	41.389$_{+0.07\%}$	7.620$_{+0.05\%}$	67.22$_{+63.95\%}$
ResNet101 + CA	46.146$_{+11.57\%}$	**7.942$_{+4.28\%}$**	41.42$_{+1.02\%}$
ResNet101 + Ours	41.610$_{+0.60\%}$	7.627$_{+0.14\%}$	**40.94$_{-0.15\%}$**

The best values under different descriptions are shown in bold

Table 6 Ablation experiment data of the channel embedded spatial module

Description	Parameters (M)	GFLOPs	CIFAR-10 Error (%)
MobileNetV2 (Baseline)	**2.237**	0.326284	17.98
+ CA	2.682$_{+19.89\%}$	0.336191$_{+3.04\%}$	17.80$_{-1.00\%}$
+ Ours(M + A)	2.682$_{+19.89\%}$	0.332205$_{+1.81\%}$	17.62$_{-2.00\%}$
+ Ours(M + A + C)	2.682$_{+19.89\%}$	**0.339166$_{+3.95\%}$**	**17.20$_{-4.33\%}$**

The best values under different descriptions are shown in bold

[13], and CA (space embedded channel attention mechanism) [17].

Table 1 clearly shows that the proposed SPCII module improves the performance of the MobileNetV2 and ResNet18 baseline networks on the CIFAR-10 and CIFAR-100 datasets, further verifying its universality on different network architectures.

With respect to the MobileNetV2 model on the CIFAR-10 dataset, the SPCII module achieves a more significant error rate reduction than do the other attention mechanism modules, such as the SE, ECA, CBAM, and CA modules. The SPCII module proposed in this paper has the lowest error rate, which is reduced by 4.33% compared to that of the baseline network. When the MobileNetV2 model is applied to the CIFAR-100 dataset and the ResNet18 model is applied to the CIFAR-10 and CIFAR-100 datasets, the SPCII module effectively reduces the error rates by 2.32%, 8.87%, and 5.72%, respectively. The excellent performance of this paper's method in terms of classification accuracy stems from the fact that the adaptability of SPCII enables the model to focus on key regions in different spatial dimensions, making it more effective at capturing the local features and location information of the targets. In addition, on the CIFAR-100 dataset, the SPCII module always performs well on the MobileNetV2 and ResNet18 models, demonstrating its robustness in handling datasets with multiple categories.

Furthermore, the impact of the SPCII module on the model parameters is also investigated. The results in Table 1 show

that the parameter size of SPCII on the ResNet18 algorithm is 11.256 M, which reduces the Parameters by 0.19% and 11.82% compared to those of the SE and ECA algorithms, respectively. Compared with ResNet18, the SPCII increases the parameter by only 0.66%. This indicates that while maintaining a relatively low parameter increase, the SPCII algorithm achieves lower error rates on the CIFAR dataset and higher GFLOPs than do the other algorithms.

Comparison between different ResNet depths

In this section, the robustness of the SPCII is demonstrated by deepening the model depth of ResNet. As shown in Table 2, the SPCII algorithm reduces the error rate by 2.68%, 7.51%, 6.57% and 8.54% compared to those of the baseline algorithms ResNet20, ResNet32, ResNet56 and ResNet101 on the CIFAR-10 data, respectively. On the CIFAR-100 data, the SPCII algorithm reduces the error rate by 3.80%, 10.33%, 3.09% and 0.63% compared to those of the baseline algorithms ResNet20, ResNet32, ResNet56, and ResNet101, respectively. Even though the model depth of ResNet increases, the error rate of SPCII still decreases and outperforms that of the other attention mechanism modules, which fully demonstrates the power of SPCII. In Fig. 3, as the ResNet depth gradually increases, the proposed SPCII algorithm achieves a relatively low error rate with a slight increase in parameter size and computational complexity, demonstrating its performance advantage in deep networks. The SPCII module introduces an adaptive channel interaction mode, which can adaptively focus on information from different positions and spatial dimensions in the input tensor. In deep networks, it can better adapt to different levels of feature representation and capture more complex features and relationships.

Image classification on the STL-10 dataset

Target classification experiments are conducted on the STL-10 dataset to evaluate the SPCII attention mechanism module, following the training rules and parameters mentioned in Sect. "Experiment Setup", and the SPCII module is embedded into the MobileNetV2 and ResNetX series of target classification networks. First, SPCII is embedded into the MobileNetV2 and ResNet18 backbone models. Then, the ResNet depth is increased to observe the robustness of the SPCII.

Comparison between different backbone models

Table 3 shows the performance variations in MobileNetV2 and ResNet18 in terms of the number of parameters, computational complexity, and classification error rate on the STL-10 dataset. In addition, the accuracy of the proposed SPCII algorithm is better than that of other attention mechanism modules, but the SPCII algorithm increases the parameter size and computational complexity of MobileNetV2.

With respect to the MobileNetV2 model on the STL-10 dataset, the SPCII module achieves a more significant improvement in the error rate by 3.91% compared to the other attention mechanism modules, such as the SE, ECA, CBAM, and CA modules. These findings show that the SPCII module has significant performance advantages when applied to lightweight networks and small training sets. A similar trend is also verified in the ResNet18 model on the STL-10 dataset, and the SPCII module achieves an error rate reduction of 10.73%, further demonstrating its effectiveness in lightweight networks. In addition, the results also show that on the STL-10 dataset, the error rate of the MobileNetV2 model increases slightly when the SE module is added, while adding the ECA module increases the error rate dramatically, demonstrating that the use of channel attention or mixed attention mechanisms is ineffective on lightweight networks with small training sets. The small datasets fail to provide sufficiently diverse samples, thus limiting the ability of the model to learn rich feature representations. The effectiveness of the embedded design in the CA and SPCII attention mechanisms is well demonstrated.

Comparison between different ResNet depths

The training set size of the STL-10 dataset is much smaller than that of the CIFAR dataset. Therefore, the model overfits on the training set, leading to poor performance on the test set. The robustness of the SPCII in the case of a small training set is demonstrated.

As shown in Fig. 4, the error rate of SPCII decreases as the model depth of ResNet increases. This trend indicates that SPCII can still effectively improve the model performance at deeper levels and be superior to other attention mechanism modules at different depths.

In Table 4, the improvements in performance of SPCII relative to the baseline algorithm ResNet (ResNet20, ResNet34, ResNet56, and ResNet110) at different depths are shown. At layer 20, compared with ResNet20, SPCII reduces the error rate by 4.84%, and at layer 101, SPCII reduces the error rate by 0.15% compared to that of ResNet110. In contrast, other attention mechanism modules may degrade the accuracy in deep networks; in particular, SE, ECA, CBAM, and CA lose 36.59%, 16.15%, 63.95%, and 1.02%, respectively, of the accuracy. By observing the error rates of layers 20 to 101, SPCII can still achieve better performance than the other algorithms. We comprehensively consider information of the SPCII in both the horizontal and vertical directions and introduce additional combination methods during the merging process to more comprehensively understand the features

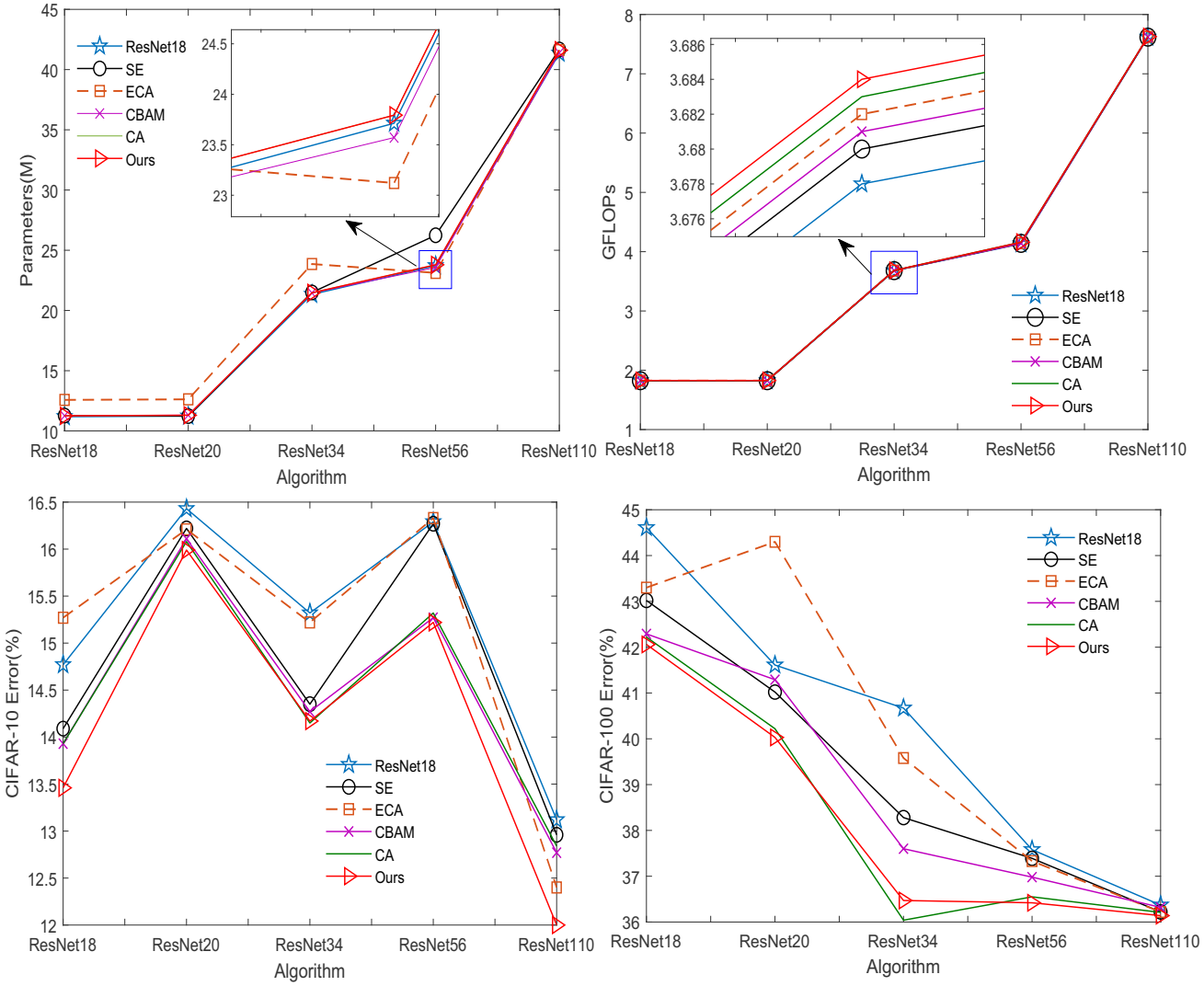

Fig. 3 Performance comparison between different attention mechanisms on different depths of the ResNet architecture on the CIFAR dataset

in the images. The SPCII is still robust when the training set is small and the network is deep.

Grad-CAM visualization result plots

To better compare the above results, Grad-CAM [34] is used to visualize the results. The thermal map generated by Grad-CAM can indicate the areas that the model focuses on in the image, thereby helping to understand how the model makes decisions. It is highly useful for local feature localization in image classification tasks. As shown in Figs. 5 and 6, the visualization results after adding SE, ECA, CBAM, CA and SPCII to MobileNetV2 and ResNet20 for comparison, on the STL-10 dataset are shown.

As shown in Fig. 5, SPCII is superior to the other algorithms in terms of coverage in the Grad-CAM visualization results of MobileNetV2. As shown in Fig. 6, SPCII is superior at capturing classification details in the Grad-CAM visualization results of ResNet20. In addition, the SPCII module is able to guide the network to focus on more important overall and detailed features while ignoring unimportant features.

For example, the third, fourth, and fifth columns in Fig. 5 are able to better focus on the head region of the target using the attention mechanism proposed in this paper, enabling the model to achieve a significant improvement in the recognition performance for specific categories. By introducing the attention mechanism, the model can enhance the representation of key regional features in a targeted manner when processing images, thus enhancing the accurate understanding towards the target.

In the third, fourth, and fifth columns of Fig. 6, using the attention mechanism proposed in this paper, more attention can be given to more detailed parts, such as the nose, eyes, and ears, and highlighting the key local information can help

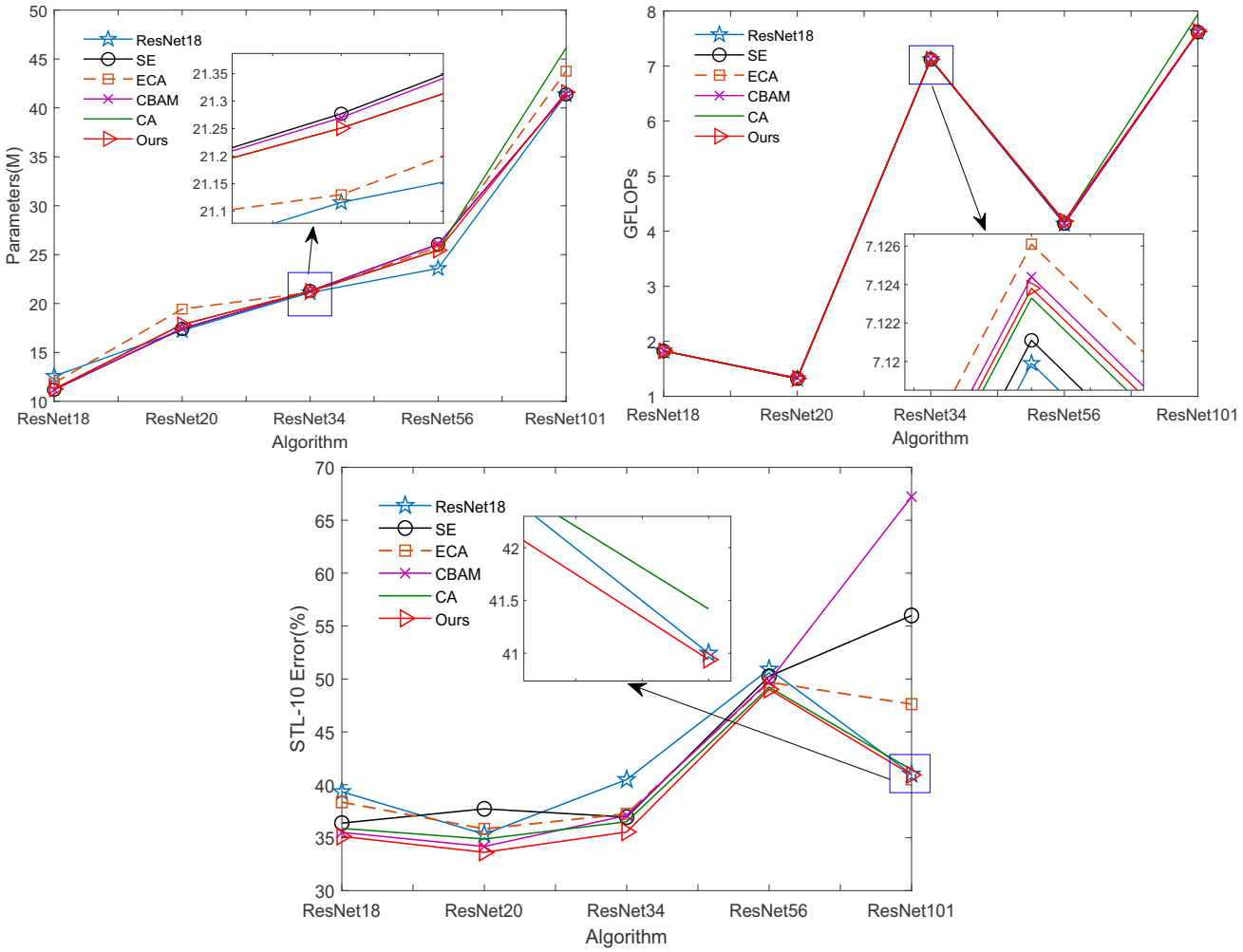

Fig. 4 Performance comparison between different attention mechanisms on different depths of the ResNet architecture on the STL-10 dataset

improve the detection accuracy of the target category in complex scenes.

Ablation studies

In the ablation experiments, the CIFAR-10 dataset is used, and MobileNetV2 is used as the backbone model. We train for 200 epochs using the parameters in Sect. "Experiment Setup" and report on the classification errors, Parameters, and GFLOPs of the test data. In Tables 5 and 6, MaxPool is abbreviated as "M", AvgPool is abbreviated as "A", and channel is abbreviated as "C".

Space embedded channel module

To validate the effectiveness of this paper's multiple pooling unit combination in the space embedded channel module, we conduct experiments on the CIFAR-10 dataset with MobileNetV2(Baseline), + CA and elimination of different pooling layers. As shown in Table 5, in the ablation experimental data of "where" embedded in "what", the algorithms MobileNetV2(Baseline) + CA, + Ours(M + C), + Ours(A + C) and + Ours(M + A + C) with the addition of the attention mechanism module reduce the error rate by 1.00%, 1.17%, 1.11% and 4.33%, respectively, when compared with the baseline algorithm. This shows that the method of combining multiple pooling units proposed in this paper is very effective at improving model performance. Compared to MobileNetV2(Baseline), the models with the addition of the attention module all achieve a significant reduction in the error rate.

Compared to MobileNetV2(Baseline), the parameters increased by 19.89% with the addition of the attention module. Compared to MobileNetV2 (Baseline), adding the CA, Ours(M + C), Ours(A + C), and Ours(M + A + C) attention modules increases the GFLOPs by 3.04% and 3.95%, respectively. Table 5 demonstrates that the method of combining

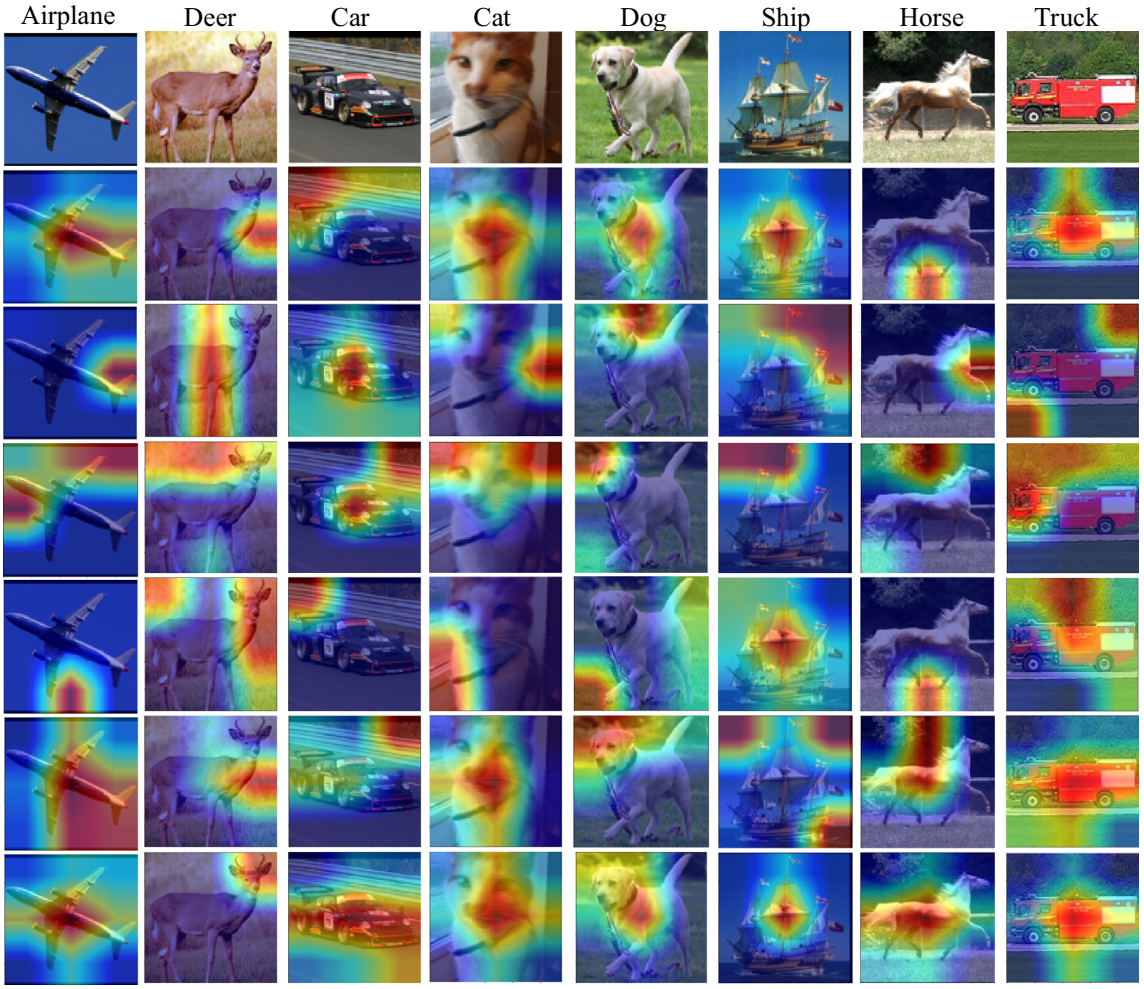

Fig. 5 Grad-CAM visualization result plots of MoblieNetV2 for partial categories on STL-10. The horizontal rows represent the original image, MobileNetV2, MobileNetV2 + SE, MobileNetV2 + ECA, MobileNetV2 + CBAM, MobileNetV2 + CA, and MobileNetV2 + SPCII images, respectively

multiple pooling units proposed in this paper is effective. Adding the attention module introduces additional additional computational overhead, but increasing the model computational power by reducing the error rate is desirable.

Channel embedded space module

To validate the effectiveness of this paper in the channel embedded space module, experiments are conducted on the CIFAR-10 dataset with MobileNetV2(Baseline) and + CA [17], and with or without the addition of the channel embedded space module. As shown in Table 6, in the ablation experimental data of "What" embedded into "Where", the algorithms MobileNetV2(Baseline) + CA, + Ours(M + A), and + Ours(M + A + C) with the addition of the attention module reduce the error rates by 1.00%, 2.00%, and 4.33%, respectively, compared with MobileNetV2(Baseline). This shows that the channel embedded space module proposed in this paper is very effective in improving the model performance. Compared to MobileNetV2(Baseline), the models of adding the attention module all achieve a significant reduction in the error rate.

Compared to MobileNetV2(Baseline), the parameters increase by 19.89% with the addition of the attention module, and the GFLOPs increase by 3.04%, 1.81% and 3.95% after adding the CA, + Ours(M + A) and + Ours(M + A + C) attention modules, respectively. The results of the ablation experiments demonstrate the effectiveness of the channel embedded space module proposed in this paper on the CIFAR-10 dataset, which is able to significantly reduce the error rate of the model. The effectiveness of the method in improving the performance of the model is demonstrated, and this method provides superior computational performance compared to other modules.

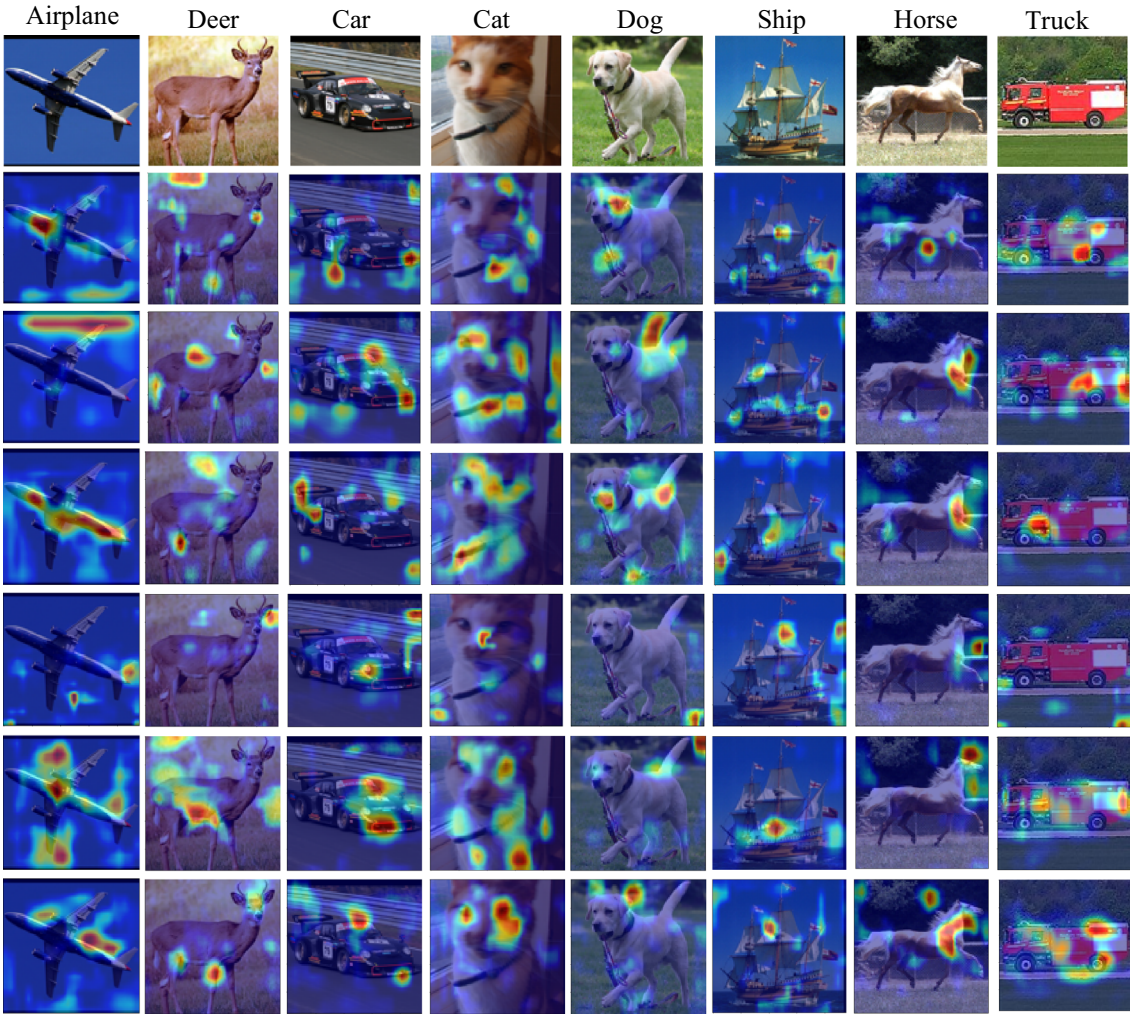

Fig. 6 Grad-CAM visualization result plots of ResNet20 for partial categories on STL-10. The horizontal rows represent the original image, MobileNetV2, MobileNetV2 + SE, MobileNetV2 + ECA, MobileNetV2 + CBAM, MobileNetV2 + CA, and MobileNetV2 + SPCII images, respectively

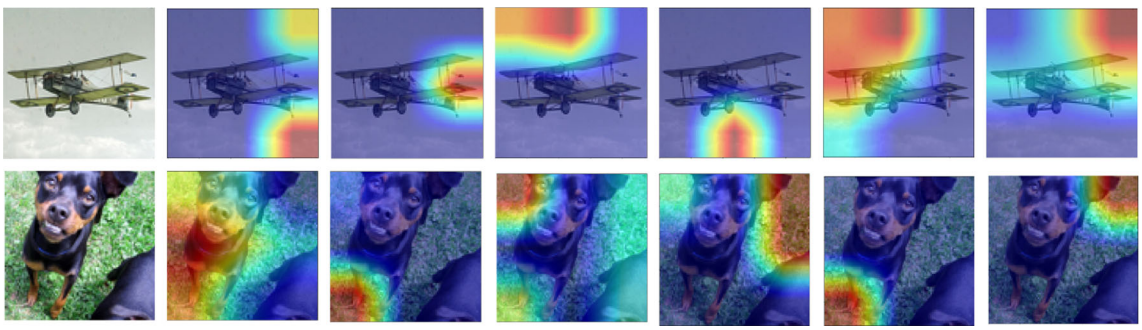

Fig. 7 Grad-CAM visualization result plots of MobileNetV2 for airplane and dog categories on STL-10. The horizontal rows represent the original image, MobileNetV2, MobileNetV2 + SE, MobileNetV2 + ECA, MobileNetV2 + CBAM, MobileNetV2 + CA, and MobileNetV2 + SPCII images, respectively

Algorithm limitations

In this section, the limitations of the algorithm proposed in this paper are analyzed through the above Grad-CAM visualization result plots. In particular, Fig. 7 shows that the algorithm cannot classify well when the target features to be classified are not obvious or when the target is occluded. Figure 7 shows the visualization results of MobileNetV2 on the STL-10 dataset with the addition of SE, ECA, CBAM, CA, and SPCII for comparison.

(1) The target features to be classified are not obvious. In the first row of Fig. 7, on the STL-10 dataset, the aircraft style is different from most styles of the aircraft category in the dataset, there is a situation where the features within the category are not obvious, and the model may fail to determine which features are critical, thus affecting its classification performance.
(2) The target to be classified is occluded. When one or more parts of the target are occluded, the model may fail to capture the shape or key features of the target completely, leading to incorrect classification. In the second row of Fig. 7, on the STL-10 dataset, the dog's ears are occluded, and the model is unable to obtain complete target information, resulting in a classification failure.

Differences from existing algorithms

This section provides a detailed analysis of representative algorithms (SE (channel attention mechanism) [4], ECA (lightweight channel attention mechanism) [5], CBAM (channel + spatial attention mechanism) [13], and CA (space embedded channel attention mechanism) [17]) in terms of logical ideas, advantages, disadvantages, and mixed approaches, and compares them with the proposed SPCII algorithm. The specific differences are as shown in Table 7.

In terms of logical ideas and mixed approaches, the SPCII algorithm differs from the SE, ECA and CBAM algorithms in that it does not solely utilize channel or spatial attention mechanisms in parallel or serially. Compared with the CA algorithm, the method proposed in this paper not only embeds the spatial dimension into the channel information, but after splitting the spatial dimension into horizontal and vertical dimensions to be embedded into the channels, the obtained channel information is embedded into the horizontal and vertical spatial dimensions, respectively, to realize the mutual embedding of the space and the channels.

In terms of advantages, compared with the SE, ECA, CBAM and CA algorithms, the SPCII algorithm not only considers the channel information in the spatial dimension, but also adopts the method of non-dimensionality reduction to realize the cross-channel information interaction, which enriches the interaction information between the spatial dimension and channel information.

In terms of disadvantages, the SPCII algorithm increases the number of parameters for network computation compared to the SE, ECA, and CA algorithms but retains local and global information, and enhances the interaction information in the remote space.

In summary, the difference between the SPCII algorithm proposed in this paper and other algorithms is that it focuses on the interaction between channel and spatial information. The SPCII algorithm embeds the spatial dimension into the channel information, enriching the channel information in the spatial dimension; the channel information is embedded into the horizontal and vertical spatial dimensions, respectively, and while considering the channel and spatial information, the cross-channel interaction of non-dimensionality reduction maintains the important relationship between the channels. Other algorithms focus more on channel information or spatial information, while the SPCII algorithm effectively integrates and interacts with the two, considering the interrelationship between channel and spatial information.

Conclusion

To improve the performance of convolutional neural network models in deep learning, this paper proposes a new attention mechanism module (SPCII) with spatial perception and channel information interaction. SPCII cascades a space embedded channel module and a channel embedded space module. The space embedded channel module, which embeds the horizontal and vertical dimensions into the channel dimension, performs global maximum and average pooling, merges the maximum and average pooling, and then splits them by the horizontal and vertical dimensions to obtain two sets of clustering features in the horizontal and vertical directions of the channels, effectively strengthening the representational competence of the object of interest. The channel embedded space module uses a channel interaction model with an adaptive convolution kernel size and embeds channel information into two spatial dimensions through 1D convolution to obtain two attention maps. In addition, ablation experiments are conducted on the CIFAR-10 dataset with MobileNetV2 as the baseline target classification network architecture, which demonstrate the effectiveness of the space embedded channel module and the channel embedded space module proposed in this paper. The SPCII is subsequently compared with popular attention modules on MobileNetV2 and various depths of ResNet architectures. The experimental results show that the proposed SPCII algorithm is optimal for improving GFLOPs and accuracy despite

Table 7 Comparison between different attention mechanism algorithms

Algorithm	Logical idea	Mixed approach	Advantages	Disadvantages
SE[4]	Squeeze and excitation	Channel attention mechanism	Enhance important channels Capture global information	Lack local information. high model complexity Long computational time
ECA[5]	Improve excitation module	Channel attention mechanism	Enhance important channels Capture global information Cross-channel interaction of non-dimensionality reduction	Lack long-distance dependencies
CBAM[13]	Predict channel and spatial attention, respectively	Channel tandem spatial attention mechanism	Focus on key regions. Establish remote dependencies Rich channel and spatial information	Overfocus on local features Increased network computation and complexity
CA[17]	Split spatial dimension into horizontal and vertical parts and embed them into channels	Space embedded channel attention mechanism	Enrich channel information in the spatial dimension, remote spatial interaction Low computational overhead	Lack local feature information Process channel information in a dimensionality reduction manner
SPCII	Split spatial dimension into horizontal and vertical parts and embed them into channels Embed channel information into spatial horizontal and vertical dimensions, respectively	Mutual embedding spatial and channel attention mechanism	Enrich the interaction information between spatial dimension and channel information Cross-channel interaction of non-dimensionality reduction Remote spatial interaction	Increase the number of network computing parameters

a slight increase in parameter size, and it has strong robustness for ResNet architectures at different depths. Finally, this paper uses Grad-CAM to perform a visual display of different attention modules on the STL-10 dataset. Finally, this paper used Grad-CAM to visualize different attention modules on the STL-10 dataset. The visualization results indicate that SPCII can more accurately focus the target classification network model on the features of the target object, achieving the real meaning of the attention mechanism. However, when the target features to be classified are not obvious or occluded, the presence of specific styles and partial occlusion of the target can affect the model performance, and future research will concentrate on improving the model's ability to adapt to these challenges. At the same time, the model will be inserted into more classification methods to verify its effectiveness on more public datasets.

Acknowledgements Thanks the Science and technology development plan of Jilin Province for help identifying collaborators for this work.

Author contribution Yifan Wang: Software, verification, writing. Yang Li: Conceptualization. Wu Wang, Yaodong Jia and Yu Xu: Format editing: Jiaqi Ma and Yu Ling Monitor.

Funding Science and technology development plan of Jilin Province (20200401090GX and 20230101174JC).

Data availability Data openly available in a public repository.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Cristina Z, Eugenio MC, Enrique HV, Iyad AK, Francisco H (2023) Explainable crowd decision making methodology guided by expert natural language opinions based on sentiment analysis with

attention-based deep learning and subgroup discovery. Inf Fusion 97(8):101821. https://doi.org/10.1016/j.inffus.2023.101821
2. Zhang S, Wei Z, Xu W, Zhang LL, Wang Y, Zhou X, Liu JY (2023) DSC-MVSNet: attention aware cost volume regularization based on depthwise separable convolution for multi-view stereo. Complex Intell 9:6953–6969. https://doi.org/10.1007/s40747-023-01106-3
3. Lakshmi RK, Rama SA (2023) Novel heuristic-based hybrid ResNeXt with recurrent neural network to handle multi class classification of sentiment analysis. Mach Learn: Sci Technol 4:015033. https://doi.org/10.1088/2632-2153/acc0d5
4. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: CVPR 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
5. Wang QL, Wu BG, Zhu PF, Li PH, Zuo WM; Hu QH (2020) ECA-Net: efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR) 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155
6. Yang ZX, Zhu LC, Wu Y, Yang Y (2020) Gated channel transformation for visual recognition. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR) 11794-11803. https://doi.org/10.1109/CVPR42600.2020.01181
7. Qin ZQ, Zhang PY, Wu F, Li X (2021) Fcanet: Frequency channel attention networks, 2021 IEEE/CVF international conference on computer vision (ICCV) 763–772, https://doi.org/10.1109/ICCV48922.2021.00082
8. Volodymyr M, Nicolas H, Alex G, Koray K (2014) Recurrent models of visual attention. Neural Inf Process Syst 2:2204–2212. https://doi.org/10.48550/arXiv.1406.6247
9. Max J, Karen S, Andrew Z, Koray Kavukcuoglu (2015) Spatial Transformer Network. NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems 2:2017–2025. https://doi.org/10.48550/arXiv.1506.02025
10. Huang ZL, Wang XG, Wei YC, Huang LC, Shi H, Liu WY, Thomas SH (2019) Ccnet Crisscross attention for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 45(6):6896–6908. https://doi.org/10.1109/TPAMI.2020.3007032
11. Park J and Sanghyun W, Lee JY, Kweon IS (2018) Bam: bottleneck attention module. ArXiv. https://doi.org/10.48550/arXiv.1807.06514
12. Li GQ, Fang Q, Zha LL, Gao X, Zheng NG (2022) HAM: Hybrid attention module in deep convolutional neural networks for image classification. Pattern Recognit J: Pattern Recognit Soc. https://doi.org/10.1016/j.patcog.2022.108785
13. Wang YB, Wang HF, Peng ZH (2021) Rice diseases detection and classification using attention based neural network and bayesian optimization. Expert Syst Appl 178:114770. https://doi.org/10.1016/j.eswa.2021.114770
14. Abhijit GR, Nassir N, Christian W (2019) Recalibrating fully convolutional networks with spatial and channel "Squeeze and Excitation" blocks. IEEE Trans Med Imaging 38(2):540–549. https://doi.org/10.1109/TMI.2018.2867261
15. Zhang QL, Yang YB (2021) Sa-net: shuffle attention for deep convolutional neural networks. ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2235–2239. https://doi.org/10.1109/ICASSP39728.2021.9414568
16. Zhang H, Zu KK, Lu J, Meng DY (2022) EPSANet: an efficient pyramid squeeze attention block on convolutional neural network. Comput Vis Pattern Recognit. https://doi.org/10.48550/arXiv.2105.14447
17. Hou QB, Zhou DQ, Feng JS (2021) Coordinate attention for efficient mobile network design. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). 13708–13717. https://doi.org/10.48550/arXiv.2103.02907
18. Le CY, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1990) Handwritten digit recognition with a back-propogation network. Adv Neural Inf Process Syst. https://doi.org/10.5555/2969830
19. Alex K, Ilya S, Geoffrey EH (2012) Imagenet classification with deep convolutional neural networks. In: 2012 neural information processing systems (NIPS) 25:1097–1105. https://doi.org/10.1145/3065386
20. Karen S, Andrew Z (2015) Very deep convolutional networks for large_scale image recognition. 2015 international conference on learning representations (ICLR). https://doi.org/10.48550/arXiv.1409.1556
21. Christian S, Sergey I, Vincent V, Alexander AA (2016). Inception-v4, inception-ResNet and the impact of residual connections on learning. AAAI'17: proceedings of the Thirty-First AAAI conference on artificial intelligence 4278–4284 https://doi.org/10.48550/arXiv.1602.07261
22. He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR) 7. https://doi.org/10.1109/CVPR.2016.90
23. Andrew GH, Zhu ML, Chen B, Dmitry K, Wang WJ, Tobias W, Andreetto M, Hartwig A (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. ArXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
24. Mark S, Andrew H, Zhu ML, Andrey Zhmoginov, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. The IEEE conference on computer vision and pattern recognition (CVPR) 4510–4520. https://doi.org/10.48550/arXiv.1801.04381
25. Andrew H, Mark S, Chu G, Chen LC, Chen B, Tan MX, Wang WJ, Zhu YK, Pang RM, Vijay V, Quoc V L, Hartwig A (2019) Searching for mobilenetv3. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.48550/arXiv.1905.02244
26. Jin HZ, Bao ZX, Chang XL, Zhang TT, Chen C (2023) Semantic segmentation of remote sensing images based on dilated convolution and spatial-channel attention mechanism. J Appl Remote Sens 17:016518–016518. https://doi.org/10.1109/LGRS.2021.3052557
27. Shen NY, Wang ZY, Li J, Gao HY, Lu W, Hu P, Feng LY (2023) Multi-organ segmentation network for abdominal CT images based on spatial attention and deformable convolution. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.118625
28. Yu Y, Zhang Y, Song Z, Tanget CK (2023) LMA: lightweight mixed-domain attention for efficient network design. Appl Intell 53(11):13432–13451. https://doi.org/10.1007/s10489-022-04170-3
29. Shen Y, Zheng W, Chen LQ, Huang F (2023) RSHAN: Image super-resolution network based on residual separation hybrid attention module. Eng Appl Artif Intell: Int J Intell Real-Time Autom 122:106072. https://doi.org/10.1016/j.engappai.2023.106072
30. Jin MX, Li HF, Xia ZQ (2023) Hybrid attention network and center-guided non-maximum suppression for occluded face detection. Multimed Tools Appl 82:15143–15170. https://doi.org/10.1007/s11042-022-13999-2
31. Shi CK, Hao YX, Li GY, Xu SY (2023) EBNAS: efficient binary network design for image classification via neural architecture search. Eng Appl Artif Intell: Int J Intell Real-Time Autom. https://doi.org/10.1016/j.engappai.2023.105845
32. Alex K (2009) Learning multiple layers of features from tiny images. Handbook of systemic autoimmune diseases 1(4). https://www.cs.toronto.edu/~kriz/cifar.html
33. Adam C, Honglak L, Andrew Y (2011) An analysis of single-layer networks in unsupervised feature learning. Int Conf Artif Intell Stat 15:215–223

34. Ramprasaath RS, Michael C, Abhishek D, Ramakrishna V, Devi P, Dhruv B (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. IEEE Int Conf Comput Vis (ICCV). 128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7