ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

A Multi-Grained Attention Residual Network for Image Classification

Wu Xiaogang[1] and Thitipong Tanprasert[2]

ABSTRACT

Attention mechanisms in deep learning can focus on critical features and ignore irrelevant details in the target task. This paper proposes a new multi-grained attention model (MGAN) to extract parts from images. The model includes a multi-grain spatial attention (MSA) mechanism and a multi-grain channel attention (MCA) mechanism. We use different convolutional branches and pooling layers to focus on the crucial information in the sample feature space and extract richer multi-grain features from the image. The model uses ResNet and Res2Net as the backbone networks to implement the image classification task. Experiments on the CIFAR10/100 and Mini-Imagenet datasets show that the proposed model MGAN can better focus on the critical information in the sample feature space, extract richer multi-grain features from the images, and significantly improve the image classification accuracy of the network.

Article information:

Keywords: Multi-grained Attention, Residual Network, Feature Extracting, Image Recognition

Article history:
Received: February 1, 2023
Revised: April 8, 2023
Accepted: May 1, 2023
Published: May 13, 2023
(Online)

DOI: 10.37936/ecti-cit.2023172.251536

1. INTRODUCTION

Deep learning is widely used in computer vision tasks, such as image classification [1], object detection [2-3], and target tracking [4]. New deep-learning methods and larger-scale neural network models have emerged to obtain better feature representation [5-6]. However, we prefer lightweight networks with computational simplicity [7-8] in many application scenarios (e.g. mobile and embedded devices) [7-8].

Computer vision tasks require the extraction of features at multiple levels and granularities. As shown in Figure 1, there are different elements in the image of the bird, from the whole to the local. Convolutional neural networks use different depths and widths to obtain features with different-sized perceptual fields [9]. Google Net [10] proposes an Inception module that fuses elements at different depth scales (Figure 2). The upper module of Inception enters three different convolutional layers and a pooling layer (3*3) with convolutional kernel sizes of 1*1, 3*3 and 5*5 to extract features at different scales. Finally, we fused convolutional 1×1 downscaling information in the multi-granularity parts. Several recent research advances in computer vision have focused on applying multi-grain feature networks [11–13].

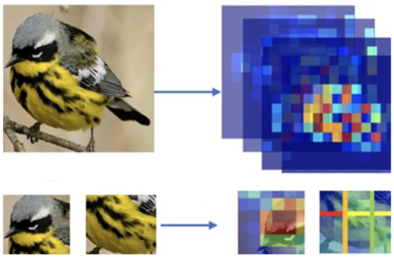

Fig.1: Multi-grained feature extraction of images.

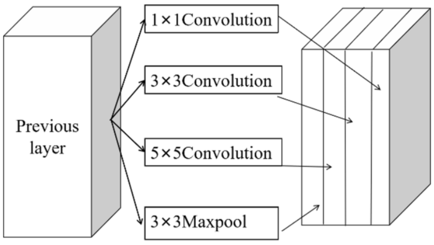

Fig.2: Multi-grained convolutional feature extraction module of Inception.

[1] The author is with the Department of Information Technology, Xingyi Normal University for Nationalities, Xingyi, China, E-mail: wxg817@163.com

[1,2] The authors are the Department of Vincent Mary School of Science and Technology, Assumption University,10240, Bangkok, Thailand, E-mail: wxg817@163.com and thitipong@scitech.au.edu

Human vision can effectively spot salient areas in complex scenes. This ability to actively select essential features from the vast amount of information is attention. Introducing attentional mechanisms to the target task enables network models to focus more on the main features and ignore irrelevant features [14-15]. While the attention mechanism enhances the feature representation of the network, its plug-and-play nature also dramatically facilitates the model's design. Using channel attention or spatial attention mechanisms can improve the effectiveness of model training [16-20]. Of the above methods, some focus only on channel[16-18] or spatial[19], and some combine information from both[20] but ignore the fusion of different granularity features. Meanwhile, Transformer with self-attention has achieved good results in computer vision [21-22]. Still, it has a high consumption of computational resources, and the performance needs to use larger datasets, while the performance suffers in the face of small-scale datasets.

Inspired by the above work, we designed a new multi-granularity attention network model (MGAN) and explored a multi-channel approach to fusing multi-granularity spatial attention. The attention model uses multi-channel pooling to combine spatial awareness at different scales, extracting multi-granularity features and constructing an end-to-end neural network training model. Experiments show that the method is general, and its classification accuracy is significantly improved.

The structure of this paper is as follows: The first part introduces the research background and ideas of this paper. Section 2 presents the multi-granularity feature extraction and attention mechanisms for residual networks. In Section 3, we give details of the implementation of the network model MGAN, and in Section 4, we provide an experimental comparison of multiple attention mechanisms for residual networks. In Section 5, we give conclusions and suggestions for further work.

2. RELATE WORK

2.1 Multi-grained Residual Module

As the depth of the convolutional network increases, problems such as overfitting and gradient disappearance often occur. The residual network ResNet proposed by Kai-Ming He [23] solves this problem well. Figure 3 shows the basic residual learning unit, x denotes the input, $F(x)$ indicates the function of residual mapping, and $H(x) = F(x) + x$ is the output of the residual unit, then the work of the deep L:

$$H(x_L) = x_l + \sum_{i=1}^{L-1} F(x_i) \tag{1}$$

$\partial Loss/(\partial x_l) = \partial Loss/(\partial H(\partial x_L))$ is the inverse gradient, where $\partial Loss/(\partial x_l)$ is the gradient descent of the loss.

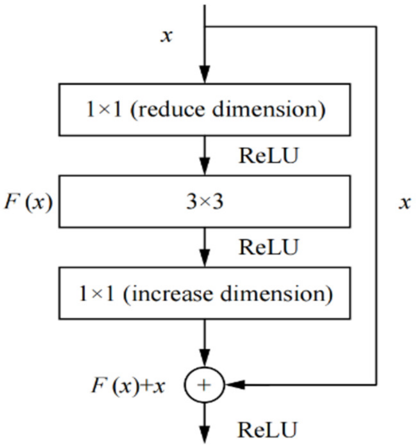

Fig.3: *The Residual unit of ResNet.*

Based on resnet,res2net [24] constructs multiple levels of residual connections within the residual block to extract multi-granularity features $f_i, i = 1, 2 \dots, s$ (Figure 4).

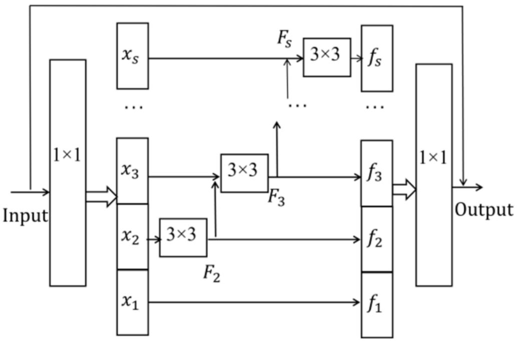

Fig.4: *The multi-Grained Residual Module of Res2Net.*

The features $f_i(i = 1, 2, \dots, s)$ obtained from the residuals of multiple branches(s denotes the number of units of the residual module), The formulated as follows:

$$f_i = \begin{cases} x_i & i = 1 \\ F_i(x_i) & i = 2 \\ F_i(x_i + f_i) & 2 < i \leq s \end{cases} \tag{2}$$

The Res2Net residual module can extract features at multiple granularities of the image through multi-level stepwise convolution, as shown in Figure 5.

2.2 Attention Mechanism

Attention mechanisms can capture more computational resources for the critical tasks of the network and reduce attention to other details when computational power is limited. The attention mechanism reflects the differentiated attention to different information through weights. When selecting the part of the input information relevant to the task, note that the importance indicates an index of that informa-

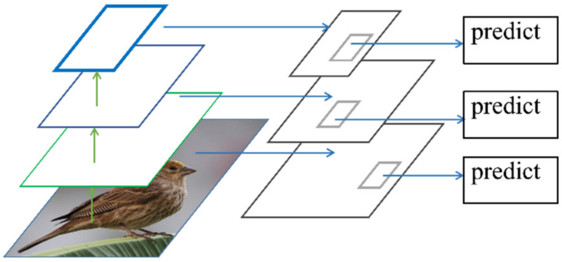

Fig.5: *Multi-granularity feature extraction of Res2Net.*

tion.[25].

Using $x_{1:N} = [x_1 \dots x_N]$ to denote N input information, $z \in [1, N]$ is the attention variable for the index position of the focal data, i.e., $z = I$ indicates the ith input message. When using soft attention [26], α_i is the probability of selecting the ith input message for the current statement: $x_{1:N}$. To extract the attention weight (3), then we have:

$$\begin{aligned}\alpha_i =& p(z = i|x_{1:N}) = softmax(S(x_i)) \\ =& \frac{\exp(S(x_i))}{\sum_{j=1}^{N} \exp(S(x_j))}\end{aligned} \quad (3)$$

Where $S(x_i)$ is the attention-scoring function, we use it to select the appropriate parameters in the model.

2.3 Feature fusion for multi-channel attention

Multi-channel attention is multi-headed attention applied to 2D image features [27]. When the model uses multiple channels of attention, the attention is a query $q_j \in \{q1, \dots, qk\}$ for multiple single-channel tasks, i.e., multiple selection processes on the input information. The multi-channel feature is a weighted summation of the single-channel components, and k is the number of channels; We have the following formula equation (4).

$$MultAttension(\alpha_i, q_j) = \sum\nolimits_{j=1}^{K} \sum\nolimits_{i=1}^{K} \alpha_i q_j \quad (4)$$

3. MULTI-GRAINED ATTENTION NETWORK

3.1 multi-grained attention module

Our proposed multi-granularity attention model comprises a multi-granularity spatial attention module MSA (Figure 6) and a multi-granularity channel attention MCA (Figure 7). Finally, these two multi-granular attentional features are fused.

The features extracted from the backbone network go through the three convolutional branches of the spatial attention module (MSA) to obtain different attentions $S^i \in F(K_i \times K_i)$ (as shown in Figure 6), where W_i and b_i denote the weights and bias parameters of the different branches, and Conv2D denotes the two-dimensional convolutional operation, we have:

$$S^i = Conv2D(I, K_i) = W_i^T + b_i, i = 1, 2, 3 \quad (5)$$

Each branch extracts the spatial attention at different granularities by maximum pooling and average pooling, respectively S^i_{max}, S^i_{avg}, $i = 1, 2, 3$. We get spatial attention weights through Concat operations, $Fs = S^i_{max} \oplus S^i_{avg}$, which $\oplus$ indicates a Concat process.

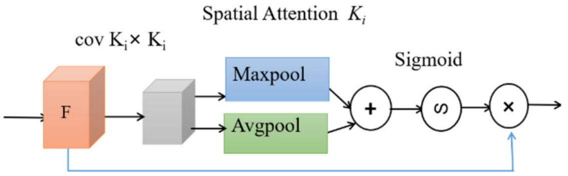

Fig.6: *The Spatial Attention with convolution kernel-K_i.*

Algorithm 1: Multi-Granularity Spatial Attention
Input: Image I and convolution kernel K_i
Output: Spatial convolution as Ki-branching attention

```
Procedure $MSA\ (I, K_i)$
$for\ each\ convolution\ K_i$ :
    $S^i = Conv2D\ (I, K_i)$
    $S^i_{max} = Spatial\_Max\_pooling(S^i)$
    $S^i_{avg} = Spatial\_Avg\_pooling(S^i)$;
    $S^i_{cat} = Sigmoid(Conv2D(concat(S^i_{max}, S^i_{avg}), 1)$
    $S^i_{cat} = S^i_{cat} \times I$
    $return\ S^i_{cat}$
$end\ for$
End Procedure
```

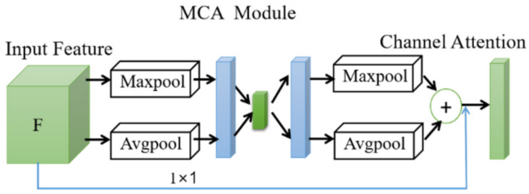

Fig.7: *Multi-Channel Attention Module.*

The spatial attention function Fs is input to the channel attention unit in the MCA structure. The maximum pooling and average pooling represent different granularity features of the image. We denote their weights by F_{max} and F_{avg}, respectively. It then enters the multilayer perceptron MLP fusion to obtain the channel attention weights Cw in equation (6).

$$C_w = Sigmoid(MLP(F_{max}) + MLO(F_{avg})) \quad (6)$$

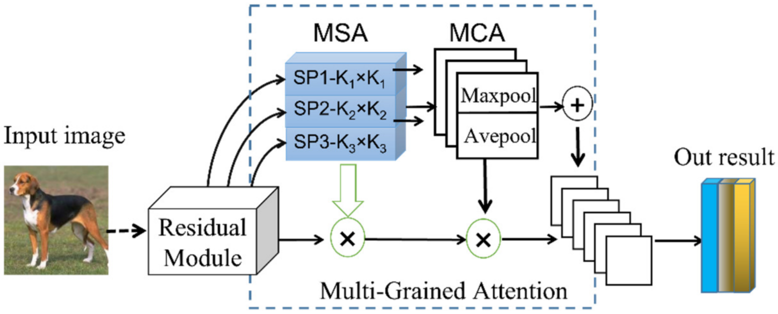

Fig.8: Image classification model for multi-granularity attention residual networks.

Where F_{max} and F_{avg} share weights in the multilayer perceptron layer. We generate a set of global features ($F_G \in R^{m\times m\times L}$) using a 1×1 convolution, then multiply F_G with the pooled feature F_{Cat} to get the multi-grain channel attention features $F_{Channel}$. The algorithm is as follows:

```
**Algorithm 2:** Multi-Grained Channel Attention
**Input:** Image $I$ and Channel $C_w$
**Output:** Channel attention
Procedure $MCA$ $(I, C_w)$
1. $Fc_{max} = Channel\_Max\_pooling(C_w)$;
2. $Fc_{avg} = Channel\_Avg\_pooling(C_w)$;
3. $MLP(Fc_{max}) = (w1(w0(Fc_{max}))$
4. $MLP(Fc_{avg}) = (w1(w0(Fc_{avg}))$
5. $F_{cat} = Sigmoid(MLP(Fc_{max}) + MLP(Fc_{avg}))$
6. $F_G = Conv1 \times 1(I))$
7. $F_{channel} = F_{cat} \times F_G$
8. $Return\ F_{channel}$
End Procedure
```

In Algorithm 3, the features extracted from the residual units of the backbone network enter different convolution branches of the spatial and channel attention modules. Then, we use the softmax function to obtain a multi-granularity vector of attention weights and feed it into the classifier network. The algorithm is as follows:

```
**Algorithm 3:** Multi-granular Attention Model
**Input:** Image $I$ and Feature of $MSA$, $MCA$
**Output:** Classification probability
Procedure MA $(I, F, K_i C_w)$
1. $S1 = MSA(I,\ K1)$
2. $S2 = MSA(I,\ K2)$
3. $S3 = MSA(I,\ K3)$
4. $Cw = MCA(I,\ Cw)$
5. $F1 = Cw(S1)$
6. $F2 = Cw(S2)$
7. $F3 = Cw(S3)$
8. $Fcat = concat(F1,\ F2,\ F3)$
9. $Fnew = concat(Fcat,\ Conv1 \times 1(F))$
10. $Fpool = Max\_pooling(Conv2D(Fnew, K))$
11. $Fc = Full\ Connection\ (flatten(Fpool))$
12. $y = softmax(Fc)$
End Procedure
```

3.2 Multi-Grained Attention Residual Network

The multi-grain attention residual network model (MGAN) in this paper consists of a residual module (ResNet/Res2Net) as the backbone, a multi-grain spatial attention module (MSA), a multi-grain channel attention module (MCA), and a classifier.

In this model, the backbone network converts the input image into underlying features; The attention algorithm reinforces the multi-granularity spatial weights of the image and combinates the consequences of the multiple channels; Then, it is fused with the underlying features of the image to obtain the multi-granularity representation of the image; and the classifier converts the high-level features into probabilities corresponding to each category in the dataset. The final output is the classification result. Figure 8 shows this network model's structure, four components, and a lightweight attentional residual network.

Multi-granularity attention mechanisms can enhance the feature representation of a network. However, overlaying attention modules can lead to degraded model performance due to the many dot product operations required for attention branches. In contrast, selecting residual networks for the leading network can reduce the decay of deep network features [28]. We use the residual module to extract features and then augment the critical components of the network with multi-granularity Attention weights (as in Figure 9).

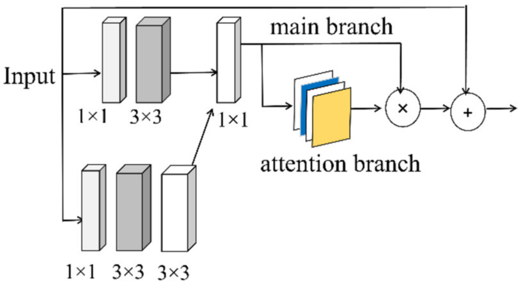

Fig.9: Structure of Multi-Attention Residual module.

Table 1: *Parameters of the backbone network ResNet/Res2Net (layers 18, 34, 50, 101).*

Layers	18-layer	34-layer	50-layer	101-layer
Conv1	7×7(3×3),64, stride=2			
	max pool 3×3, stride=2			
Conv2x	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 2$ Attention(K_i=3)	$\begin{bmatrix} 3\times3, & 64 \\ 3\times3, & 64 \end{bmatrix} \times 3$ Attention(K_i=3)	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$ Attention(K_i=3)	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$ Attention(K_i=3)
Conv3x	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 2$ Attention(K_i=5)	$\begin{bmatrix} 3\times3, & 128 \\ 3\times3, & 128 \end{bmatrix} \times 4$ Attention(K_i=5)	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$ Attention(K_i=5)	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 3$ Attention(K_i=5)
Conv4x	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 2$ Attention(K_i=7)	$\begin{bmatrix} 3\times3, & 256 \\ 3\times3, & 256 \end{bmatrix} \times 6$ Attention(K_i=7)	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 6$ Attention(K_i=7)	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 23$ Attention(K_i=7)
Conv5x	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, & 512 \\ 3\times3, & 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 23$
	Average pool, 10/100-d fc, softmax			
FLOPs	1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9

4. EXPERIMENT

The following experiments evaluate the image classification performance of our Multi-Granularity Attention Network (MGAN) with different backbone networks and parameters.

4.1 Experimental Parameters and Data Sets

Our experiments used the deep learning framework PyTorch 1.12 (Python 3.9), an NVIDIA GeForce RTX3060 graphics card with 12 GB of RAM and CUDA (version 11.2). We used ResNet and Res2Net as the backbone networks and set the size of the three convolutional kernels in the multi-grain spatial attention module to K_1=3, K_2 =5, and K_3 =7. We set the batch size to 64, the initial learning rate to 0.01, mileage nodes divided by ten every 50 training sessions, and weight decay to 0.0001. The network hyperparameters are shown in Table 1 below, where 18, 34, 50, and 101 layers are used for the backbone network, respectively.

The convolution size of conv1 is 7×7 or 3×3, depending on the size of the images in the dataset.

We evaluated our model on the datasets CIFAR-10, CIFAR-100[29], and Mini-ImageNet[30]. The dataset CIFAR-10 contains 60,000 32 × 32-pixel colour images, 50,000 training images and 10,000 test images, divided into ten categories. (Figure 10), and the CIFAR-100 dataset contains 100 types of 600 pictures each. CIFAR-100 has 100 classes divided into 20 broad classes, so each image has a coarse and fine label. The Mini-ImageNet dataset is a small sample dataset extracted by the Google DeepMind team based on ImageNet. The dataset contains 60,000 colour images 84 × 84 pixels in size, divided into 100 categories with 600 samples in each category.

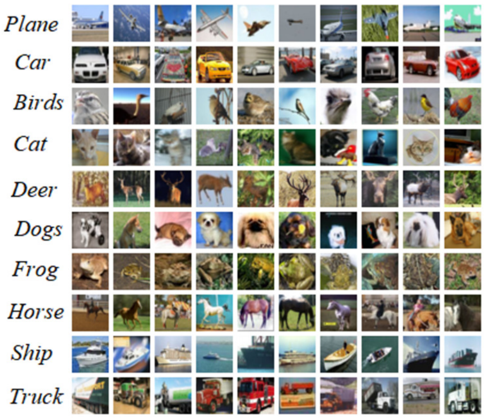

Fig.10: *Example image of dataset Cifar10.*

The performance evaluation metrics we use include Top_1 accuracy (%), Top_5 accuracy (%), training loss, and training time (hours).

We define Accuracy as the probability of correct classification of the test set as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

Top_1 Accuracy is the probability that the maximum predicted category is the correct category, while Top_5 Accuracy is the Accuracy of giving the top 5 indicated categories that contain the proper category.

4.2 Compare the Result with a different attention network

In the first part of the experiments, we compare the MGAN network with other attention networks, such as SENET [31], CBAM [32], and SimAM[33], on the three different datasets.

ResNet18 is the backbone network, allowing easy comparison with other commonly used attention networks. Figure 11, Figure 12, and Figure 13 show the Top1-Acuraccy and Top5-Acuraccy classification performance of the different attention models on the three datasets, respectively.

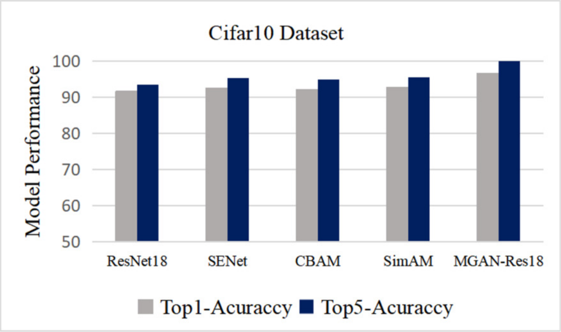

Fig.11: Results on the dataset Cifar10.

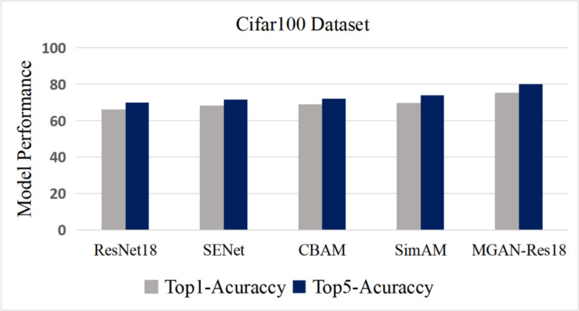

Fig.12: Results on the dataset Cifar100.

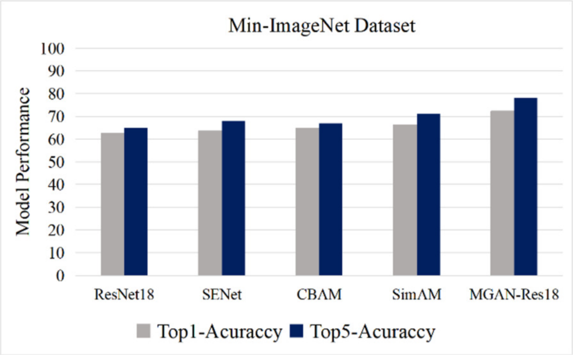

Fig.13: Results on the dataset Mini-ImageNet.

The experimental results in Table 2 show that our model has higher classification performance than other attention models on the same backbone, ResNet18.

In practice, we will use the more powerful backbone network Res2Net50[24], fusing the extracted multi-scale features with a multi-grain attention mechanism to obtain a richer quality. In the following experiments, we choose several networks with multi-scale attention [34-35] and compare them on the dataset Min-imageNet, where our model performs slightly better when the parameters are similar (as shown in Table 3 and Figure 14).

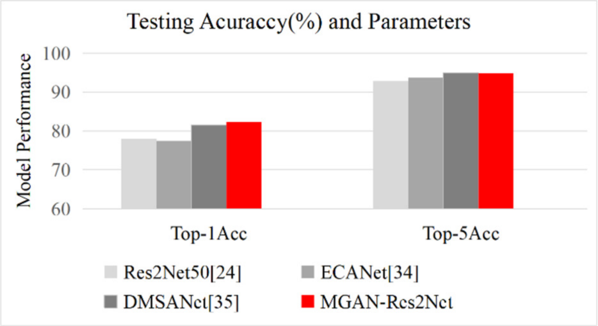

Fig.14: Performance comparison of the multi-granularity attention networks.

4.3 Compare the Performance of Different Backbones

The second part of the experiments compares the performance of different residual modules with the multi-grained attention model. We first chose ResNet18 as the backbone of the multi-granularity attention mechanism (MGAN-ResNet18) and trained the dataset Cifar10 with epoch=100. Figure 15 shows that our MGAN network quickly achieves over 90% training accuracy at epoch=20 on dataset CIFAR10 and can reach 96.5% accuracy on the test set at epoch=100. Meanwhile, the value of the loss also decreases rapidly. In this experiment, we specify the learning rate at epoch=50 to one-tenth of the original, i.e., 0.001; We note that the loss value at this point drops significantly (see Figure 16).

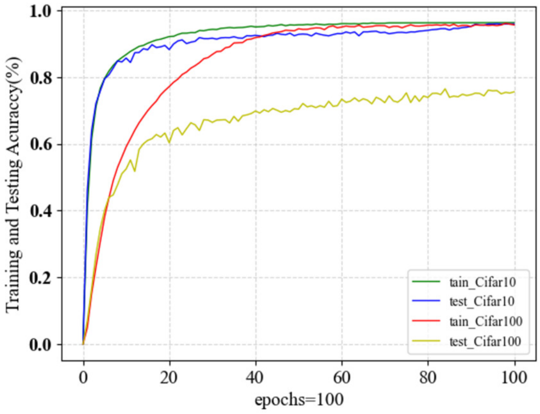

Fig.15: Accuracy of the network MGAN-ResNet18 on Datasets Cifar10.

We then chose ResNet and Res2Net with different layers as the backbone network to test our network performance on the CIFAR100 dataset. We set epoch=100 and select the number of layers x to be

Table 2: *Top1 Testing accuracies (%) and Top5 Testing accuracies (%)for the backbone networks: ResNet18 with various attention modules and the proposed MGAN.*

Models	Params	CIFAR10		CIFAR100		Mini-ImageNet	
	10^6	Top1(%)	Top5(%)	Top1(%)	Top5(%)	Top1(%)	Top5(%)
ResNet18[23]	11.69	91.50±0.50	93.5.0±0.50	66.0±0.36	70.05±0.45	62.40±0.30	65.82±0.35
+SE[31]	11.78	92.42±0.14	95.20±0.40	68.70±0.21	71.20±0.35	63.59±0.30	68.32±0.40
+CBAM[32]	11.78	92.19±0.11	94.90±0.45	68.76±0.56	72.16±0.20	64.83±0.30	67.45±0.50
+SimAM[33]	11.69	92.73±0.18	95.50±0.50	69.57±0.40	74.25±0.40	66.31±0.30	71.25±0.40
MGAN-ResNet18	11.70	96.50±0.40	99.90±0.10	76.55±0.50	80.10±0.50	72.26±0.50	78.30±0.50

Table 3: *Comparing different multi-granularity attention mechanisms on the dataset Mini-ImageNet.*

Models	Params	Top-1Acc	Top-5Acc
	10^6	%	%
Res2Net50[24]	25.56	77.99	92.87
ECANet[34]	25.56	77.48	93.68
DMSANet[35]	26.25	81.54	94.93
MGAN-Res2Net	28.6	82.40	94.85

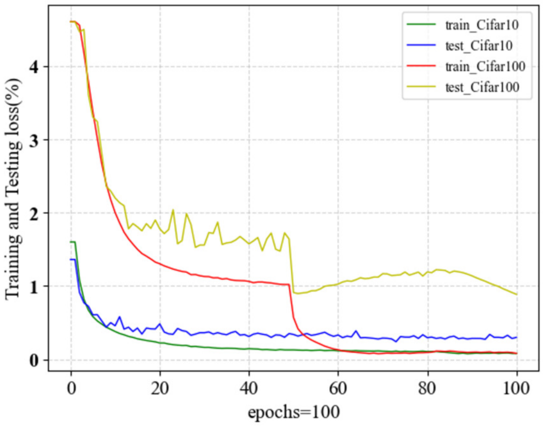

Fig.16: *Accuracy of the model MGAN-ResNet18 on Datasets Cifar100.*

18,34,50,101; Figure 17 shows the classification accuracy of the test set on different Backbone: ResNetx and Res2Netx; Figure 18 and Figure 19 show the training time on the three datasets.

Figure 17 shows that the classification accuracy of the network gradually improves as the number of layers of the backbone network increases, and of course, the training time increases. And the overall performance of using Res2Net as the backbone network is more robust than that of ResNet.

4.4 Comparing the performance of spatial attention modules with different parameters K_i

We compared the effects of different spatial attention parameters on the model. In Table 1, we set the three convolutional kernel sizes in the spatial attention module to K1 = 3, K2 = 5, and K3 = 7, respectively, and connect the same channels of attention. Considering that we used datasets with image sizes of 32×32 and 84×84 pixels, respectively, We use

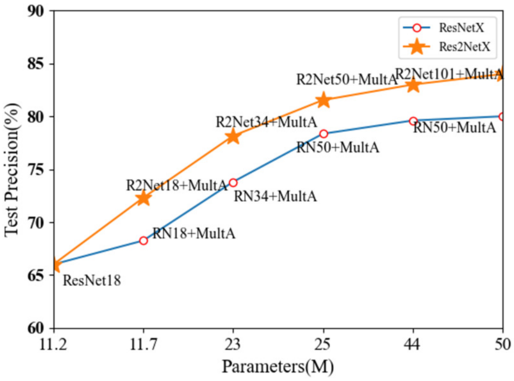

Fig.17: *Testing Precision on dataset Cifar100 with different Backbone ResNet and Res2Net.*

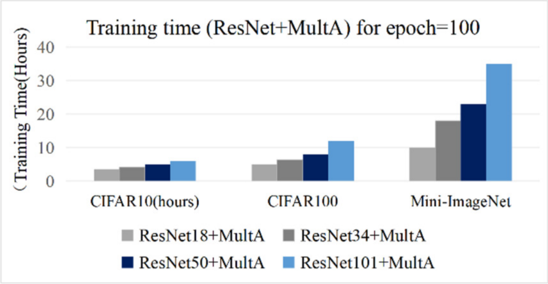

Fig.18: *Training time for different backbone ResNetx with mult-grained attention model.*

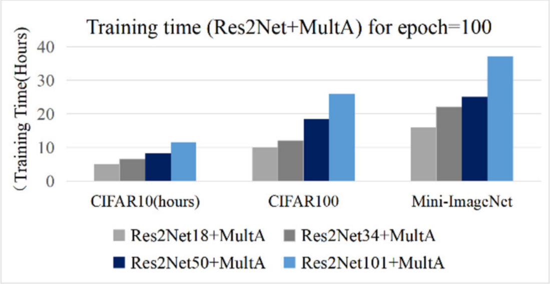

Fig.19: *Training time for different backbone Res2Netx with mult-grained attention model.*

these three small convolutional kernels in our experiments: 3 × 3,5 × 5 and 7 × 7. In contrast, for datasets with image pixels larger than 224 × 224, we prefer large convolutional kernels of 11 or more (e.g. 31 × 31) [36]. We compared the performance of four combined models with different spatial attention parameters on three datasets with a backbone using Res2Net18. The experimental parameters and results are shown in Figure 20 and Table 4.

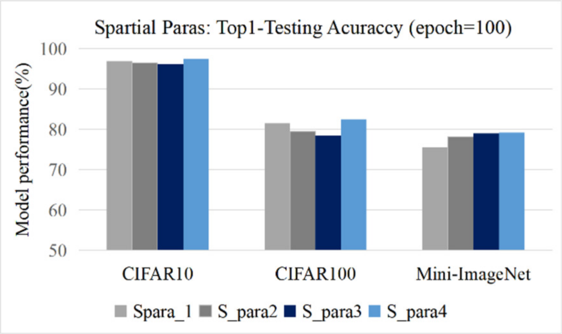

Fig.20: *Comparison Results of different parameters K_i.*

Table 4: *Performance (%) of various parameters K_i .*

Models	Spatial paras	CIFAR 10	CIFAR 100	Mini-ImageNet
S-para1	K1=K2=K3=3	96.9	81.5	75.5
S-para2	K1=K2=K3=5	96.5	79.5	78.2
S-para3	K1=K2=K3=7	96.2	78.5	79.0
S-para4	K1=3,K2=5,K3=7	97.5	82.50	79.20

Figure 20 shows that the performance using the same convolution parameters is slightly lower than that of the combined model S-Para4 using different convolutions. In the S-Para1, S-Para2, and S-Para3 models with the same size convolution, the Cifar10 and Cifar100 datasets are more suitable with convolution kernels of 3*3 and 5*5. In contrast, the Mini-ImagaNet dataset is better suited to a 7*7 convolutional kernel, mainly due to the size of the images in the corresponding dataset.

4.5 Comparison of the performance of different types of networks

In this part of the experiment, we compared the testing accuracy of the networks with varying parameter scales with our model to evaluate the performance (Table 5).

As can be seen from Table 4 and Figure 21, our multi-granularity attention network has higher classification accuracy than the other three networks with the smallest parameter size. MGAN-Res2Net improved over the VGG16 network by 4%, 9%, and 7% on the three datasets with only one-fifth of its parameters.

Table 5: *Top1 Testing the Accuracy (%) of different networks with various parameters .*

Network	Param 10^6	CIFAR10	CIFAR100	Mini-ImageNet
VGG-16[10]	138	93.50 ±0.06	72.95 ±0.04	72.40 ±0.04
Res2Net 18[24]	23.2	95.50 ±0.40	75.60 ±0.50	74.54 ±0.50
WideRes Net [37]	36.5	96.11 ±0.40	79.50 ±0.50	75.50 ±0.50
MGAN-Res2Net	28.2	97.50 ±0.50	82.50 ±0.40	79.20 ±0.50

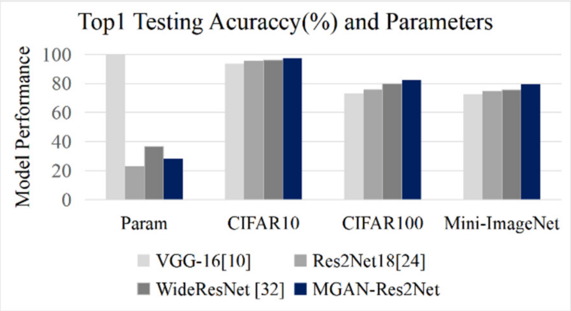

Fig.21: *Comparison Results of different parameter networks.*

The above experiments showed that combining our multi-granularity attention mechanism with various residual networks significantly enhanced classification performance.

5. CONCLUSIONS

We propose in this paper a multi-granularity attention mechanism as a new approach to improve the representational power of residual networks, which can obtain richer high-level features of images and enhance image classification accuracy. The proposed multi-granularity attention mechanism lets neural networks focus on important feature information at multiple levels in a snap. This approach transforms the original way neural networks allocate resources equally to an image to assign weights to different spatial groups of the importance of an image, thus enabling faster and more accurate classification of images.

We conducted extensive experiments on the Multi-Granularity Attention Network (MGAN) using networks with different parameters. We showed significant improvements in classification performance on the datasets Cifar10, Cifar100, and Mini-ImageNet, thus validating the effectiveness of the multi-granularity attention model.

In the following work, we will further investigate the introduction of multi-granularity methods into self-attentive networks such as Transforms [19], applying this approach to various fields such as fine-grained image classification and target detection.

ACKNOWLEDGMENT

This work was supported by the Science Foundation of Minzu Normal University of Xingyi (No.20XYJS01).

References

[1] B. Cheng *et al.*, “Revisiting rcnn: On awakening the classification power of faster rcnn,” *Computer Vision – ECCV 2018: 15th European Conference*, pp.473-490, 2018.

[2] C.-Y. Wang, A. Bochkovskiy and H.-Y. Mark Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” *arXiv preprint arXiv:2207.02696*, 2022.

[3] S. Mekruksavanich and A. Jitpattanakul, “FallNeXt: A Deep Residual Model based on Multi-Branch Aggregation for Sensor-based Fall Detection,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 16, no. 4, pp.352-364, 2022.

[4] I. Ahmed and G. Jeon, “A real-time person tracking system based on SiamMask network for intelligent video surveillance,” *Journal of Real-Time Image Processing*, vol. 18, pp. 1803-1814, 2021.

[5] H. M. D. Kabir *et al.*, “SpinalNet: Deep Neural Network With Gradual Input,” in *IEEE Transactions on Artificial Intelligence*, 2022.

[6] K. Han *et al.*, “Transformer in transformer,” *Advances in Neural Information Processing Systems*, vol. 34, pp.15908-15919, 2021.

[7] Q. Yan *et al.*, “A Lightweight Network for High Dynamic Range Imaging,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, pp. 823-831, 2022.

[8] Iandola, Forrest N., et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size.” arXiv preprint arXiv:1602.07360 (2016).

[9] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings*, Part I 13. Springer International Publishing, 2014.

[10] C. Szegedy *et al.*, “Going deeper with convolutions,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1-9, 2015.

[11] T.-Y. Lin *et al.*, “Feature pyramid networks for object detection,” *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 936-944, 2017.

[12] X. Wang *et al.*, “Towards multi-grained explainability for graph neural networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp.18446-18458, 2021.

[13] Y. Li *et al.*, “Mvitv2: Improved multi-scale vision transformers for classification and detection,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794-4804, 2022.

[14] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao and H. Lu, “Attention CoupleNet: Fully Convolutional Attention Coupling Network for Object Detection,” in *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 113-126, Jan. 2019.

[15] J. Zhang, L. Niu and L. Zhang, “Person Re-Identification With Reinforced Attribute Attention Selection,” in *IEEE Transactions on Image Processing*, vol. 30, pp. 603-616, 2021.

[16] H. Zheng, J. Fu, T. Mei and J. Luo, “Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition,” *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 5219-5227, 2017.

[17] W. Li *et al.*, “Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification,” *Neurocomputing*, vol. 387, pp. 63-77, 2020.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, “ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 11531-11539, 2020.

[19] X. Li, W. Wang, X. Hu and J. Yang, “Selective Kernel Networks,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 510-519, 2019.

[20] Y. Hu, J. Li, Y. Huang and X. Gao, “Channel-wise and spatial feature modulation network for single image super-resolution,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3911-3927, 2019.

[21] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[22] K. Han ,A. Xiao, E. Wu *et al.*, “Transformer in transforme,” *Advances in Neural Information Processing Systems*, vol. 34, pp.15908-15919, 2021.

[23] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp.770-778, 2016.

[24] S. -H. Gao, M. -M. Cheng, K. Zhao, X. -Y. Zhang, M. -H. Yang and P. Torr, “Res2Net: A New Multi-Scale Backbone Architecture,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652-662, 2021.

[25] A. Borji and L. Itti, “State-of-the-Art in Visual

Attention Modeling," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, Jan. 2013.

[26] Z. Li *et al.*, "Seq2seq dependency parsing," *Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

[27] Li, Jian, et al. "Multi-head attention with disagreement regularization," *arXiv preprint arXiv:1810.10183*, 2018.

[28] F. Wang *et al.*, "Residual attention network for image classification," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6450-6458, 2017.

[29] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[30] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 4353-4361, 2015.

[31] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132-7141, 2018.

[32] S. Woo *et al.*, "Cbam: Convolutional block attention module," *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[33] L. Yang *et al.*, "Simam: A simple, parameter-free attention module for convolutional neural networks," *International conference on PMLR*, 2021.

[34] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 11531-11539, 2020.

[35] A. Sagar, "Dmsanet: Dual multi scale attention network," *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings*,Part I. Cham: Springer International Publishing, pp. 633-645, 2022.

[36] X. Ding *et al.*, "Scaling up your kernels to 31×31: Revisiting large kernel design in cnns," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[37] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

Wu Xiaogang received his Master's degree in applied mathematics from the School of Information Science, Guangzhou University in China. He is now Studying for his Ph.D. in Computer Science at the Faculty of Vincent Mary School of Science and Technology, Assumption University, Bangkok, Thailand. He is currently a Faculty Member with the Department of Computer Science, Xingyi Normal University in China, His current research interests include Artificial Intelligence Applications and computer vision.

Thitipong Tanprasert is an Assistant Professor in Computer Science at Vincent Mary School of Science and Technology, Assumption University of Thailand, where he is also directing the Intelligent Data Analytics Research Laboratory. He received his bachelor degree in Electrical Engineering from Chulalongkorn University in 1987, the master and doctor of philosophy in Computer Engineering degrees from University of Louisiana at Lafayette in 1989 and 1993, respectively. His research interests are in neural network computation, data science, and machine learning applications.