IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 26, 2021, accepted March 8, 2021, date of publication March 10, 2021, date of current version March 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065338

LADNet: An Ultra-Lightweight and Efficient Dilated Residual Network With Light-Attention Module

JUNYAN YANG, JIE JIANG, YUJIE FANG, AND JIAHAO SUN
College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
Corresponding author: Jie Jiang (jiejiang@nudt.edu.cn)

ABSTRACT Image classification task is an important branch of computer vision. At present, most of the mainstream CNNs are large in size and take up too much computing resources. The quality-price ratio is not satisfying when the heavy CNNs are used in image classification. So, this work proposes a spatial and channel hybrid attention module (Light-Attention module), an ultra-lightweight but efficient attention module. Given an intermediate feature map, the Light-Attention module will firstly derive the most important attention maps of the channels automatically with global stochastic pooling and Multilayer Perceptron (MLP). Then, the residual structure of Light-Attention module helps to repeatedly introduce global information for secondary screening. For better performance, our pioneering Max module and Mean module to extract key spatial features on the basis of previous operations. In the process of extracting the key attention map, LADNet uses the most labor-saving operation and save a lot of network parameters. The Light-Attention module can be seamlessly integrated into any network, while its parameters and required computing resources are both very small. We verified the excellent performance of the Light-Attention module in extracting key image information through ablation experiments on the Cifar-10, Cifar-100 and ImageNet datasets. For the experiment in Cifar-10, the image classification accuracy achieves 98.7%, achieves 96.5% for the Cifar-100, and for the ImageNet, achieves 83.9%. With the Light-Attention module, LADNet reduces the parameters of convolution neural network to 71% of the original. Compared to the other CNNs used in image classification task, the Light-Attention module uses the least parameters but scores most advanced achievements. The experiments proved that the combination of DRN (Dilated Residual Network) and Light-Attention module achieves superior performance.

INDEX TERMS Light-Attention module, global stochastic pooling, DRN (Dilated Residual Network).

I. INTRODUCTION

As an important subject in the field of computer vision, image classification has always been a hot topic for scientists. Image classification, as the name suggests, is a problem of inputting images and outputting descriptions of the image content classification. As one of the earliest research topics in the field of computer vision, image classification plays an important role in image segmentation and object detection in the domain of computer vision. In recent years, the field of image classification develops so rapidly, among which attention mechanism and convolution neural network are particularly outstanding.

The associate editor coordinating the review of this manuscript and approving it for publication was Genoveffa Tortora.

With its powerful ability of image feature extraction, attention convolutional neural network has been widely used for image segmentation, image classification and other tasks in the field of computer vision. Generally speaking, there are two important parts in the attention convolutional neural network: Backbones and Attention Mechanism.

Firstly, Backbones:

From the LENet [1], which was originally applied for bank handwritten signature recognition, to the ResNet [2], which is particularly popular now, convolutional neural network architecture has been continuously deepened and broadened. VGGNet [3] deepens the network by replacing the large convolution kernels with the small convolution kernels, which verifies that the network performance can be improved by deepening the network depth. GoogleNet [4] proposes

the Inception Module to solve the problem of increasing parameters caused by network deepening, and also verifies that widening the network can reduce the difficulty of network training and increase the complexity of network model. In the process of deepening and widening the network, researchers have found some problems: when the depth of the network increases to a certain extent, the accuracy of network training decreases, and at the same time, the Gradient vanishing problem happens. Therefore, Kaiming He proposed ResNet, which enables data to flow across layers in the network through identity mapping. While ensuring that network parameters and computational complexity are not increased, the initial input is directly and leapingly introduced to the deeper network layer, so that the whole structure converges towards the direction of identity mapping. ResNet with shortcut connections successfully solved the gradient dispersion problem caused by the deepening of the network, and extended the network depth to 1000 layers. ResNet, as an easy to train and effective network, is widely used as a backbone in the later stage.

On the basis of ResNet, we consider to improve it by adopting dilated convolution [5] , in order to achieve a better backbone network in image classification tasks. Compared with the standard convolution, the dilated convolution can enlarge the receptive field without losing information by pooling, so that each convolution output contains a large range of information.

Secondly, Attention Mechanism:

In principle, the attention mechanism is to extract and strengthen the vital feature information of the task in the image through a certain algorithm, while weakening and suppressing the useless feature information that the task is not interested in. By processing the feature information of different importance, attention mechanism can improve the flow efficiency of feature information and reduce the redundancy of calculation. Since image feature information is divided into spatial feature information and channel feature information, attention mechanism can be divided into spatial attention mechanism, channel attention mechanism and spatial channel mixed attention mechanism. At present, most of the existing attention mechanisms have 3 common problems [6] (which are listed below):

1) The extensive calculation;
2) When the amount of data is small, it is easy to over fit;
3) The rough segmentation of feature maps and the lack of additional information lead to inaccurate results of attention allocation.

This work proposes a new attention module — Light-Attention module. For improving the way of extracting important feature information in attention mechanism, the most efficient structure that can extract effective features is selected to reduce network parameters. At the same time, additional image information is introduced to guide attention allocation.

On the Cifar-10, Cifar-100 and ImageNet datasets, the Light-Attention module is inserted into each baseline network to improve the accuracy of image classification and verify the effectiveness of the Light-Attention module. By visualizing the output feature map of each convolution layer and comparing it with the output feature map of baseline network, it can be seen that this module enables the baseline network to capture the effective feature information related to the task more quickly and accurately. Experiments show that the Light-Attention module makes the network more lightweight, optimizes the network's ability to extract effective features, and improves the network performance. In the process of designing the model, the network structure and parameters are excluded with less effect and lower cost performance ratio, so this model can be better applied to most devices.

Our main contributions are as follows:

1) Building a smaller and better attention allocation module — Light-Attention module, which can be embedded in most mainstream networks;
2) Building a DRN which can expand the receptive field and capture the global information perfectly;
3) Through ablation experiments, we prove that the effectiveness of the Light-Attention module;
4) Carrying out a list of experiments proved that the image classification accuracy of LADNet is improved by 1.2% in Cifar-10, by 3.916% in Cifar-100, by 1.9% in ImageNet (compared to the most advanced attention classification network).

II. RELATED WORK

A. ATTENTION MECHANISM

There is a selective attention principle in the human visual perception system. Due to the limited computing resources available to the human brain, when the image acquired by the eyes is transmitted to the cerebral cortex, the brain will discard the invalid information in the image and pay attention to the important and partial information related to our purpose. According to the mechanism of human visual attention, attention module has been proposed to allocate attention.

STN [7] proposed by Google DeepMind carries out pre-processing operations suitable for tasks by learning the deformation of input feature map, strengthens the expression of effective feature information in image space, and weakens the sense of existence of invalid information. In SENet [8], the attention model uses the squeeze and citation operations to model the importance of each feature channel, and enhances or suppresses different channels for different tasks.

As the most important reference of this work, the CBAM [9] model proposed by SangHyun woo combines spatial attention and channel attention in a serial way in the feedforward convolutional neural network attention module, and the attention maps are multiplied to the input feature map for adaptive feature refinement. After studying the structure, we think that the combination of spatial attention and channel attention has parameter waste, which is easy to over fit when the amount of data is small. At the same time, it does not introduce additional parameters into the attention module, so it

can't extract all the effective feature information completely and accurately.

For building Light-Attention module, we filter various existing attention module structures to retain the most accurate structure for effective feature extraction. At the same time, residual path is introduced into the attention module to help the attention module extract effective features more completely by using additional information. Through the experimental verification, Light-Attention module is proved to be able to extract important image information more efficiently and accurately. In the Cifar-10 and Cifar-100 data sets, the baseline network embedded with Light-Attention module performs better in image classification tasks, and the convergence speed of the network is three times faster than that of the original network, the network parameters are greatly reduced, and the requirements for computing resources are not high, and the practicability is strong.

B. DILATED CONVOLUTION

Convolution is a complex function widely used in neural networks to extract image features. If the attention mechanism is used to extract the important features required by the task, then the convolution extracts the features and attributes of the image according to certain logic. Convolution uses a filter (convolution kernel) to filter the small regions of the image, so as to obtain the features of them. In practical applications, researchers often use multiple convolution kernels. Each convolution kernel represents an image mode.

With the development of deep learning, many kinds of convolutions, such as expanding convolution, deformable convolution, deep separable convolution and transposition convolution, have been proposed. Because the common convolution neural network has the following problems:

1) Up-sampling is deterministic;
2) Loss of internal data structure and spatial hierarchical information;
3) Small object information cannot be reconstructed.

Some scholars have proposed dilated convolution to reduce the image size and enlarge the receptive field without increasing the network parameters.

We apply the dilated convolution to ResNet to replace the original standard convolution. By using the dilated convolution to expand the receptive field, the data flowing between each convolution layer in the ResNet is more complete, and the ResNet can extract more effective information. At the same time, dilated convolution can replace part of pooling operation to further reduce the amount of network parameters and computational complexity.

III. OUR APPROACH

A. OVERVIEW

This work proposes an image classification network–LADNet, which is very lightweight but has good performance in image classification. There are two key parts in LADNet: Light-Attention module and Dilated ResNet. The Light-Attention module is designed on the basis of the current mainstream attention mechanism. By embedding global stochastic pooling which is proposed for the first time in this work into the prototype network, LADNet greatly reduces the network parameters and the overfitting problem. The combination of global stochastic pooling and Multilayer Perceptron (MLP) [10] successfully screened out the key feature channels. At the same time, the Max module and Mean module are designed to extract the important regions on the feature map. In order to make full use of the excellent features of Light-Attention module, the DRN is proposed as the backbone of the Light-Attention module. Instead of standard convolution, LADNet uses the dilated convolution to expand the receptive field of convolution layer and extract more complete effective features. The combination of Light-Attention module and DRN greatly improves the accuracy of image classification task and reduces the network parameters.

B. LIGHT-ATTENTION MODULE

In recent years, some researchers have proposed attention mechanism by imitating the attention function of human brain when reading pictures, which has been widely used in image processing tasks. In the field of computer vision, attention mechanism mainly includes channel attention mechanism and spatial attention mechanism. The importance degree of each characteristic channel and local spatial region is modeled and calculated by using the weighted idea. Deep learning is used to enhance the effective channel and local spatial region, and suppress the channel and local space region with invalid or weak information.

In Light-Attention module, we take an intermediate feature map $F \in R^{C\times H\times W}$ in the convolution layer as the input. After being processed by Light-Attention module, a three-dimensional spatial and channel hybrid attention map $M \in R^{C\times H\times W}$ is outputted, the whole process is expressed by the formula below:

$$M = f_M(F) \tag{1}$$

where $f_M()$ means the function, which summarizes the whole algorithm of Light-Attention. The specific process and detailed description of the Light-Attention module in extracting mixed attention of spatial and channels are shown in Figure 1.

1) GLOBAL STOCHASTIC POOLING

In the design of attention mechanism, people often need to use Max pooling and Average pooling to increase the receptive field, maintain the image shift invariance, and reduce the difficulty and parameters of networks' optimization. Pooling is the process of abstracting the information of the feature maps. Max pooling and Average pooling are used repeatedly in the attention module of the latest convolution neural networks, such as DANet [11], CBAM, and SENet. Although pooling itself does not increase the number of network parameters, a large number of feature maps produced by pooling will impose a certain burden on the network during the later process of feature extraction. Therefore, Light-Attention module

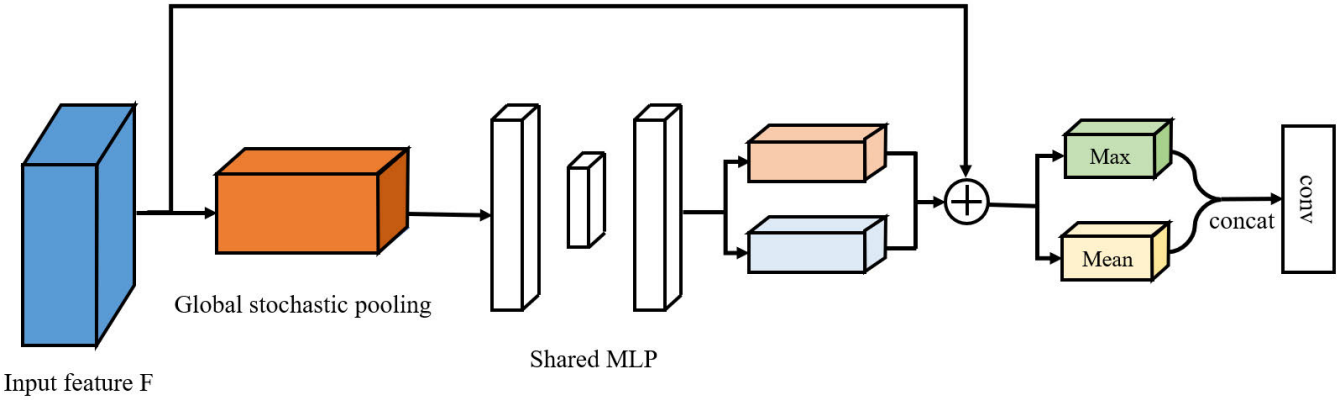


FIGURE 1. Light-Attention module: The feature map is input into Global Stochastic pooling and shared MLP to extract the key channel features, and two channel feature maps are output. The two channel feature maps and the original input feature map are added by element. Then the Max module for extracting texture information and the Mean module for extracting image background information get the added feature map as the input.

replaces the Max pooling and Average pooling with Global Stochastic Pooling.

Similar to the principle of Max pooling, Stochastic Pooling [12] randomly selects elements in a feature map according to their probability values. The probability of elements being selected is positively related to their numerical values. Stochastic Pooling retains the advantages of Max pooling and Average pooling, which can selectively preserve images' background information and highlight images' texture information. It likes an operation of regularization which enhances generalization ability of pooling.

We list the mathematical formulas of Stochastic Pooling process below:

$$A^* = \frac{A}{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}} \tag{2}$$

$$a = rand(A^*) \times (\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}) \tag{3}$$

where A is a n-dimensional matrix and means the inputted feature map (pooling stride is n), A^* means the probability matrix obtained by normalizing the feature map A; $rand()$ means the function that takes a value randomly; a is the pooled value.

In the backward propagation derivation, this structure only needs to keep the value of the selected node recorded by forward propagation, and sets other values to 0.

Inspired by the global average pool, we designed the Global Stochastic Pooling which applies the Stochastic Pooling to the whole image, (Originally, Stochastic Pooling is only applied to local areas of the image), for extracting channel features.

The schematic diagram is shown in Figure 2.

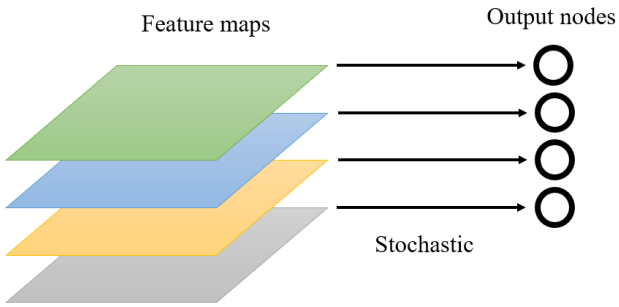


FIGURE 2. Global Stochastic Pooling.

2) MAX MODULE AND MEAN MODULE

After extracting channel features, we pay our attention to the spatial features.

In the design of spatial attention mechanism, the extraction of image background features and image texture features should be balanced. In general, Max pooling can reduce the deviation of estimated mean caused by convolution parameter error, which can be used to extract texture information of image; Average pooling can reduce the increase of estimated value variance caused by neighborhood size limitation, which is used to extract the background information of image. However, both Max pooling and Average pooling have the problem of losing image detail information and interrupting the process of back-propagation. Therefore, we consult the principle of Max pooling and Average pooling to design modules that can not only extract image background and texture information, but also retain image detail information, as well as ensure the continuity of gradient feedback. In the Light-Attention module of this paper, we propose the Max module and Mean module to meet the above requirements.

In the Max module, the mean value of the matrix is used to expand the influence of the larger value in the matrix and inhibit the role of the small and medium value in the matrix, for highlighting the expression of image texture information. In the Mean module, for consideration of reducing the amount of network parameters, the mean value of the matrix with another algorithm which expands the influence of the value close to the mean value in the matrix is still used to weakens

the action of the extreme value in the matrix, for retaining more background information of the image.

$$mv = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}}{n^2} \tag{4}$$

$$A_{ij}^{max} = \left(\frac{A_{ij}}{mv}\right)^{k} \tag{5}$$

$$A_{ij}^{mean} = \frac{A_{ij}}{|A_{ij} - mv|} \tag{6}$$

where mv means the mean value of the matrix; A_{ij} is the value of input feature map A, which is a n-dimensional matrix; A_{ij}^{max} means values of the output feature map A^{max} from Max module; A_{ij}^{mean} means values of output feature map A^{mean} from Mean module;

3) PARAMETER REDUCTION

In Light-Attention module, the Global Stochastic Pooling which is designed on the basis of Stochastic-pooling and Multilayer Perceptron (MLP) are used to extract important feature information in images' channels. Considering that the traditional attention extraction mostly uses the combination of mean pooling and maximum pooling, which requires convolution neural network to calculate a large number of parameters, resulting in a waste of computing resources (Although pooling itself will not increase network parameters, a large number of feature maps generated by multiple pooling will indirectly cause the burden of network training in the later MLP calculation). Global Stochastic pooling is used in Light-attention module to produce only one feature map a time, which greatly reduce Computing resources token by MLP in the next step. It assigns probability to pixels according to numerical value, and then subsamples according to probability. In average sense, it is similar to mean pooling, while in local sense, it obeys the principle of maximum pooling. Stochastic-pooling can not only better complete the function of channel feature information extraction and integration, but also greatly reduce the amount of network parameters. The feature map obtained by stochastic pooling is input into MLP, and the channel features with higher importance are transformed by using activation function, which can compensate certain information in deep convolution.

In order to further introduce additional information, we refer to the principle of residual network, and introduce residual edge path in Light-Attention module. The attention feature maps from MLP are elementwise added with input feature map F. The input feature map F is reused for secondary information filtering to gain the most intact and effective image information at the minimum cost, so as to reduce the burden of network training.

After the channel feature extraction, Max module and Mean module respectively extract texture information and background information from the image by using mean value of the matrix and our pioneering algorithm. What's more, it's worth mentioning that the Max module and Mean module do not introduce other parameters, so it will not burden network training, and can ensure the continuity of back-propagation, which makes up for the traditional pooling Insufficient.

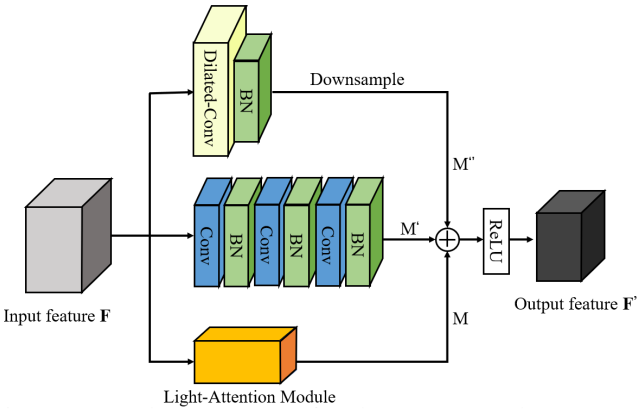


FIGURE 3. Dilated Residual Network with Light-Attention module.

C. DRN (DILATED RESIDUAL NETWORK)

As shown in Figure 3, the main line of Dilated Residual network is a stack of convolution and batch normalization. In the residual path, we use Dilated Convolution [5] to replace the traditional standard convolution. Ensuring a certain amount of network parameters, it expands the receptive field so that the output of each convolution contains a larger range of information, which helps Light-Attention module to extract important features to a certain extent. The main line, down sampling and Light-Attention module are in parallel. These three kinds of output feature maps M, M ', M' are added and input into the ReLU function together, then we get the output feature map.

Layers stacking can deepen the depth of the network without the problem of gradient vanishing. At the same time, the training difficulty of the whole network is greatly reduced.

IV. EXPERIMENT

The experiment is divided into two parts: Ablation Experiment, which is used to verify the effectiveness of Light-Attention module; Image Classification Experiment, which is used to verify the performance of LADNet in image classification on Cifar-10, Cifar-100 and ImageNet datasets.

A. EXPERIMENT DETAILS

1) CIFAR-10

Ablation experiments and image classification experiments are done on Cifar-10 dataset. The Cifar-10 dataset consists of 60000 color images with size of 32∗32 from 10 classes, and each class has 6000 images. There are 50000 training images and 10000 test images.

2) CIFAR-100

Image classification experiments are done on Cifar-100 dataset. Cifar-100 is like Cifar-10, but it has 100 classes, each class contains 600 images. Each class has 500 training images and 100 test images. The 100 classes in Cifar-100 are divided into 20 superclasses. Each image has a ''fine'' tag (the class it belongs to) and a ''coarse'' tag (the superclass it belongs to).

3) IMAGENET

The image classification experiments are carried out on ImageNet dataset. The ImageNet dataset contains 14, 197, 122 images and 21, 841 synsets. ImageNet dataset has always been the benchmark for evaluating the performance of image classification algorithms.

4) TRAINING DETAILS

In the ablation experiment, we test the network and output the feature maps for comparison. In the image classification experiment, we adopt the Cifar and ImageNet datasets. For the Cifar dataset, we use a batch size of 2048 to train 500 epochs. The initial learning rate is 0.1 which is reduced by 0.005 for every 40 epochs. For the ImageNet dataset, we use 128 batch size to train 700 epochs, the initial learning rate is 0.3 which is reduced by 0.005 for every 45 epochs. The pretrained model of ImageNet on ResNet-50. (Due to the limitation of computer resources, only two CPUs are used in all the experiments in this paper, and all the experimental data are obtained by the author from the experiment) is pre-adopted.

B. ABLATION EXPERIMENT

In this stage, we set ResNet-50 [2] as the basic architecture, and CBAM module, which is more advanced in the field of attention mechanism, as the control group. The ResNet-50 embedded with Light-Attention module and ResNet-50 embedded with CBAM module were trained on the dataset, and the output feature maps of each convolution layer was visualized. The experimental results of the two groups were compared.

By comparing the feature maps of the Light-Attention module and the CBAM module embedded in ResNet-50, it can be observed that 75% of the feature maps from convolution layer embedded with Light-Attention module capture the key information completely, while 31.25% of the feature maps from convolution layer embedded with CBAM module capture the key information completely. (the standard of capturing key information is from user research based on Grad Cam Visualization [13]. Defining the feature map with gradient cam value greater than or equal to 0.6 can capture complete key information.)

Through the comparison of the effect of key information capture, we verify that Light-Attention module is better than CBAM module in the performance of key information capture, and indirectly surpasses most of the current attention modules. The Light-Attention module has universal applicability in different architectures.

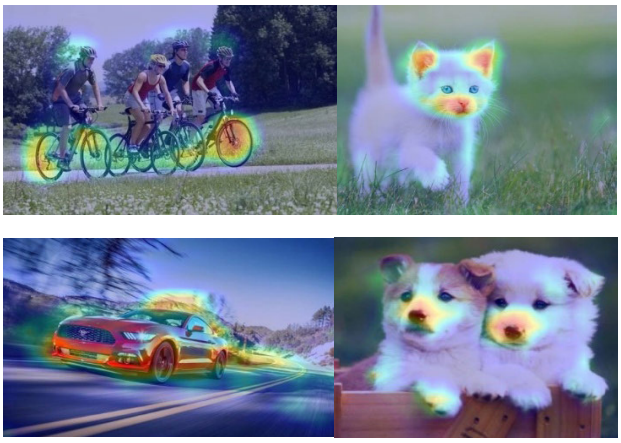


FIGURE 4. Output heatmaps of Light-Attention module.

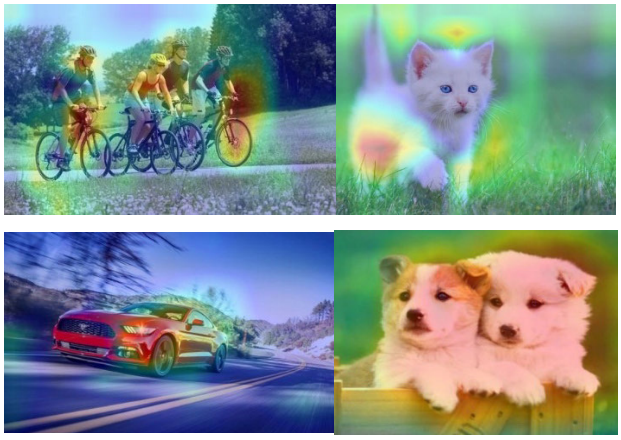


FIGURE 5. Output *heat*maps of CBAM module.

C. IMAGE CLASSIFICATION EXPERIMENT ON CIFAR-10, CIFAR-100 AND IMAGENET DATASETS

1) COMPARISON OF DIFFERENT COMBINATIONS' PERFORMANCE IN IMAGE CLASSIFICATION

Based on the Cifar-10 and Cifar-100 datasets, the effect of the DRN (Dilated Residual Network) embedded with the Light-Attention module for image classification is verified. In the second experiment stage, we embed the Light-Attention module into the mainstream baseline for experiments.

In the current image classification field, CBAM and BAM represents the most advanced image classification algorithm with attention module which were born in 2018. In order to highlight the superior performance of LADNet, we use CBAM model for comparative experiments. At the same time, we also use ResNet-50, SENet and other mainstream convolutional neural networks as prototype networks, for proving the excellent ability of LADNet to reduce the amount

of network parameters as well as improve the image classification accuracy.

Finally, in order to verify that the performance of LADNet is better than that of the network which also focuses on reducing network parameters and improving network performance, we select LGAM which is the latest lightweight attention CNN to do the last comparative experiments.

In Table 1, we summarize the results of image classification experiments on Cifar-10 and Cifar-100 using mainstream convolutional neural networks and LADNet. Data show that LADNet can achieve higher accuracy than other networks on Cifar dataset.

TABLE 1. Comparison of image classification performance in CIFAR-10 and CIFAR-100 datasets.

Model	Cifar-10 acc.%	Cifar-100 acc.%	Epoch
CBAM	97.5	92.58	50
BAM [15]	98	93	50
SENet	97	91.5	50
ResNet-50	97	91	55
LGAM [16]	95.3	91	55
WRN [17]	96.7	90.3	50
LADNet	**98.7**	**96.5**	**40**

As Table 2, 3, 4, 5 show, on the Cifar-10 and Cifar-100 datasets, we choose SE, CBAM and Light-Attention modules to embed in three network architectures which are ResNet-50, ResNeXt [17] and DenseNet [18], for image classification experiments. Table 2, 3, 4, 5 have summarized the experimental results of this stage. The data show that the performance of Light-Attention modules' performance for the image classification task is better than SE module and CBAM module. Meanwhile, the combination of Light-Attention modules and Dilated Residual Module can achieve the best performance. This means that the residual structure of stochastic pooling, Max and Mean module is powerful, which can better integrate image information and grasp key points.

we find that the fitting speed of DRN with LA modules in the Cifar-10 and Cifar-100 datasets is about 1.5 times faster than that of general network, and its total cost in terms of network parameters and consumption of computing resource shows a good advantage.

In order to illustrate the excellent performance of LADNet in image classification better, we carried out experiments again on the ImageNet dataset which is more authoritative. The experimental data show that LADNet can obtain better classification accuracy than the mainstream neural networks in the field of image classification (including the attention CNN), and the convergence speed of LADNet is faster.

TABLE 2. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the DRN.

Module +	CBAM	SE	**LA**
Cifar-10 acc.%	98.3	98.35	**98.7**
Cifar-100 acc.%	94.9	94.3	**96.5**
EPOCH	40	50	**40**

TABLE 3. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the ResNet-50.

Module +	CBAM	SE	**LA**
Cifar-10 acc.%	98.3	98.2	**98.52**
Cifar-100 acc.%	93	93	**93.2**
EPOCH	40	55	**40**

TABLE 4. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the Resnext.

Module +	CBAM	SE	**LA**
Cifar-10 acc.%	98.13	98.10	**98.65**
Cifar-100 acc.%	95.1	94.7	**95.4**
EPOCH	50	55	**45**

TABLE 5. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the DenseNet.

Module +	CBAM	SE	**LA**
Cifar-10 acc.%	98.67	98.39	**98.74**
Cifar-100 acc.%	95.69	94.8	**96.23**
EPOCH	50	60	**50**

TABLE 6. Comparison of image classification performance.

Model	TOP-1 acc.%	TOP-5 acc.%	Epoch
CBAM	82	96.7.0	80
BAM [15]	82.5	96.0	80
SENet	82	96.2	80
ResNet-50	76	93.0	90
LGAM [16]	78	94.0	90
WRN [17]	79	94.0	80
LADNet	**83.9**	**97.2**	**60**

TABLE 7. Networks' parameters and flops.

MODEL	PARAM.	GFLOPS
CBAM	28.1	3.86
BAM	23.68	3.37
SENET	25.8	3.87
RESNET-50	25.56	3.86
LGAM	23.46	3.98
WRN	23.52	3.88
LADNET	**20.4**	**4.21**

TABLE 8. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the DRN.

MODULE +	CBAM	SE	**LA**
PARAM.	28.2	25.9	**20.4**
GFLOPS	3.87	3.88	**4.21**

TABLE 9. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the ResNet-50.

MODULE +	CBAM	SE	**LA**
PARAM.	28.2	26.1	**21.5**
GFLOPS	3.86	3.86	**3.96**

TABLE 10. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the Resnext.

MODULE +	CBAM	SE	**LA**
PARAM.	28.3	26.7	**21.1**
GFLOPS	4.2	4.11	**4.1**

TABLE 11. Comparison of the effects of SE module, CBAM module and light-attention module embedded in the Densenet.

MODULE +	CBAM	SE	**LA**
PARAM.	29.3	28.2	**22.3**
GFLOPS	4.12	3.96	**4**

2) COMPARISON OF NETWORKS' PARAMETERS AND FLOPS

From table 7, we can see that the parameters are far less than the most superior CBAM module and SE module.

The parameters and GFLOPs of the Network embedded with CBAM module, SE module and Light-Attention module are calculated and analyzed by comparative experiments. The data in Table 8, 9, 10, 11 show that among different attention modules, Light-Attention module has the lightest burden on the Network. The Network embedded with Light-Attention module generally improves the training efficiency and achieves higher image classification performance.

Through the above experimental data analysis, the superiority of our network structure can be well proved. In order to simplify the structure of the model to the greatest extent, we choose the structure which is most helpful to the performance of attention extraction. In order to reduce the occupation of computing resources, LADNet directly discard some invalid information in some stages of network training, and use this manipulation to reduce the amount of network

parameters and reduce the training difficulty. We embed CBAM module, SE module and Light-Attention module into the Dilated Residual Network, ResNet-50, ResNeXt and DenseNet which are used for image classification experiments on Cifar-10, Cifar-100 and ImageNet datasets. The experimental data show that the Dilated Residual Network embedded with the Light-Attention module can obtain higher image classification accuracy than other attention modules and mainstream convolution neural networks. The parameters of the network are less, and it is easier to train.

It can be said that Light-Attention module is the first super ultra-lightweight attention module at present. Its performance on Cifar-10, Cifar-100 and ImageNet datasets is generally better than other combinations. Analyzing experiments' results, we suggest embedding the Light-Attention modules in DRN, which can not only greatly improve the accuracy of image classification task, but also maximize network performance, it has great potential in low-end devices and took up fewer computing resources.

V. CONCLUSION

We propose a new ultra-lightweight convolutional neural network attention module (Light-Attention modules), which can reduce the consumption of parameters and computing resources, and ensure high image classification accuracy. Using all kinds of pooling [19] flexibly, our attention module adopts the method of first collecting channel attention, then collecting spatial attention based on the cumulative effect of previous operation, which greatly reduces the redundancy of operation, obtains better attention than CBAM module and SE module, and processes the effective information in the image more completely. At the same time, we design a dilated residual network to combine with Light-Attention module. Through the image classification experiments on Cifar-10 and Cifar-100 datasets, we prove that the Light-Attention module can greatly improve the performance of different network architectures, and greatly accelerate the speed of network training. Combined with the visualization results of the output feature maps of convolutional layer in Ablation Experiment, we prove that the feature extraction ability of Light-Attention module is better than most of the existing attention modules, especially the dynamic saliency extraction [20] and Laurent Itti's static saliency extraction [21]. It is the first ultra-lightweight attention module which doesn't sacrifice computing resource to get better attention extraction ability. We hope that the Light-Attention module can be widely used in various network architectures and get improved in different image processing tasks.

VI. FUTURE SCOPE

At present, the development of image classification is extremely rapid, and the existing mainstream neural network has been able to achieve very good classification accuracy on the existing datasets. However, the existing scientific research achievements for the practical application of social life is far from human requirements. We need to further improve the network training speed, training accuracy and network occupied computing resources. At the same time, most of the data to be processed are disorderly and unlabeled data, which puts forward higher requirements for unsupervised or weakly supervised image classification tasks.

Therefore, in the future research, we will further study the lightweight image classification, and apply the results of this paper in the later unsupervised image classification task to further improve the effect of unsupervised image classification. Based on meta learning and reinforcement learning, we plan to design a greater model to be applied to unsupervised learning in the next step, combined with the module proposed in this paper.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[4] K. S. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[5] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.

[6] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," 2019, *arXiv:1904.02874*. [Online]. Available: http://arxiv.org/abs/1904.02874

[7] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2016.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[10] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, Aug. 1998.

[11] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2018, *arXiv:1809.02983*. [Online]. Available: http://arxiv.org/abs/1809.02983

[12] S. Zhai, H. Wu, A. Kumar, Y. Cheng, Y. Lu, Z. Zhang, and R. Feris, "S3Pool: Pooling with stochastic spatial sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4970–4978.

[13] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vision.*, vol. 128, 2019, pp. 618–626.

[14] J. Park, S. Woo, J. Y. Lee, and I. Kweon, "BAM: Bottleneck attention module," in *Proc. ECCV*, 2018.

[15] P. Zhang, Q. Li, and C. Yang, "Image classification algorithm based on lightweight group-wise attention module," *Comput. Appl.*, vol. 40, pp. 645–650, Apr. 2019.

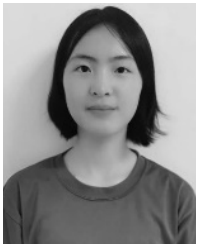
[16] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016.

[17] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

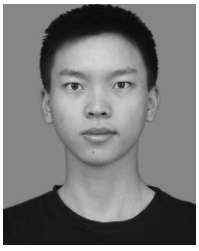
[20] Z. Jie and W. Wei, "Saliency extraction based on visual attention model," *Comput. Technol. Develop.*, vol. 20, no. 11, pp. 109–113, 2010.
[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Dec. 1998.



JUNYAN YANG received the B.S. degree from the National University of Defense Technology, in 2020, where she is currently pursuing the master's degree in visualization with the College of System Engineering. Her research interests include deep learning and computer vision.

JIE JIANG received the Ph.D. degree in control science and engineering from the National University of Defense Technology, China, in 2010. He is currently an Associate Professor and the Head of the Teaching and Research Section with the College of Systems Engineering, National University of Defense Technology. He has authored over 40 academic papers in peer-reviewed international journals and conferences. His research interests include computer vision, virtual reality, visualization, and visual analysis.

YUJIE FANG received the B.S. degree from Anhui University, in 2018. He is currently pursuing the master's degree in visualization with the College of System Engineering, National University of Defense Technology. His research interests include visualization and visual analysis.



JIAHAO SUN received the B.S. degree from the National University of Defense Technology, in 2019, where he is currently pursuing the master's degree in visualization with the College of System Engineering. His research interests include deep learning and computer vision.