frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 30 September 2022
DOI 10.3389/fnbot.2022.1007939

Check for updates

OPEN ACCESS

EDITED BY
Peng Li,
Dalian University of Technology, China

REVIEWED BY
Hang Li,
Shenyang Normal University, China
Jianing Zhang,
Dalian University of Technology, China

*CORRESPONDENCE
Chengdong Wu
wuchengdongnu@163.com

RECEIVED 31 July 2022
ACCEPTED 26 August 2022
PUBLISHED 30 September 2022

CITATION
Zhang J, Yu X, Lei X and Wu C (2022) A novel CapsNet neural network based on MobileNetV2 structure for robot image classification.
Front. Neurorobot. 16:1007939.
doi: 10.3389/fnbot.2022.1007939

COPYRIGHT
© 2022 Zhang, Yu, Lei and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A novel CapsNet neural network based on MobileNetV2 structure for robot image classification

Jingsi Zhang, Xiaosheng Yu, Xiaoliang Lei and Chengdong Wu*

Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China

Image classification indicates that it classifies the images into a certain category according to the information in the image. Therefore, extracting image feature information is an important research content in image classification. Traditional image classification mainly uses machine learning methods to extract features. With the continuous development of deep learning, various deep learning algorithms are gradually applied to image classification. However, traditional deep learning-based image classification methods have low classification efficiency and long convergence time. The training networks are prone to over-fitting. In this paper, we present a novel CapsNet neural network based on the MobileNetV2 structure for robot image classification. Aiming at the problem that the lightweight network will sacrifice classification accuracy, the MobileNetV2 is taken as the base network architecture. CapsNet is improved by optimizing the dynamic routing algorithm to generate the feature graph. The attention module is introduced to increase the weight of the saliency feature graph learned by the convolutional layer to improve its classification accuracy. The parallel input of spatial information and channel information reduces the computation and complexity of network. Finally, the experiments are carried out in CIFAR-100 dataset. The results show that the proposed model is superior to other robot image classification models in terms of classification accuracy and robustness.

KEYWORDS

robot image classification, CapsNet neural network, MobileNetV2, attention module, spatial and channel information

Introduction

In recent years, convolutional neural networks (CNN) have developed rapidly and achieved fruitful results. In terms of military application, convolutional neural network is used to identify and detect military targets (Choi et al., 2020; Yin and Li, 2020; Zeng et al., 2021). However, in the environment of missile-borne terminal, there are high requirements on network and hardware, which requires the network to maintain lightweight and good embedding performance on the basis of ensuring recognition accuracy. After AlexNet achieves a qualitative leap in the classification accuracy of ImageNet large data set (Prabhu et al., 2020), the development trend of convolutional neural network is complicated and the number of convolutional layers is increasing.

Representative networks include VGG (Simonyan and Zisserman, 2014) and GoogLeNet (Szegedy et al., 2015), which respectively, study the depth and width of convolutional neural networks and increase the network depth and width to improve the network performance. In 2015, He et al. (2016) innovatively proposed the ResNet network, which introduced the concept of residual for the first time and used residual to transmit information, effectively alleviating the problems of over-fitting and gradient disappearance caused by the increase of network depth. This network also provides ideas for the subsequent development of lightweight. Then, the researchers introduced the attention mechanism into the convolutional neural network. SENet (Hu et al., 2020) in 2017 used the attention mechanism and gave different weights to different feature graphs to improve the learning ability of the network. The later proposed attention modules draw on this idea and improve the SENet.

For example, CBAM module studies spatial attention on the basis of channel attention (Woo et al., 2018). The introduction of this idea makes network design tend to be simple, fast and portable. On the basis of point-by-point convolution and deep convolution, the typical lightweight network MobileNetV2 introduces the backward residual structure and linear bottleneck structure (Sandler et al., 2018), which makes the network structure smaller and the speed further be improved. Branch and Carvalho (2021) combined the MobileNetV2 network model with the traditional Hough transform. It performed better for the identification task, which could improve the accuracy rate with a lower number of parameters and calculation (Shafiq et al., 2021). On the basis of MobileNetV2 network, Yang et al. (2011) drew on the idea of DenseNet dense connection and utilized the benefits of feature reuse to improve network performance and reduce network scale (Akay et al., 2021). Hui et al. (2018) modified the long module on the basis of the bottleneck layer of MobileNetV2, which further reduced the complexity and computation amount when the network accuracy was improved. Cao et al. (2021) chose MobileNetV2 network as the backbone network of target detection and introduced channel attention mechanism to realize feature enhancement and effectively improve the detection performance of the network while ensuring the lightness of the algorithm (Jisi and Yin, 2021). The above methods have the following problems: a lot of network parameters, slow convergence time, easily falling into over-fitting. The CapsNet neural network can save computation time and improve classification accuracy.

Based on MobileNetV2 lightweight network and combined with CapsNet neural network, a channel space dual collaborative attention module is designed in this paper. By giving more weights to the extracted features, the recognition accuracy is increased while ensuring the network lightweight. On the CIFAR-100 standard data set, the effectiveness of the network is verified by comparing different typical convolutional neural networks and different attention modules.

The main structures of this paper are as follows: section Related works displays two networks including MobileNetV2 network and attention mechanism. Section Proposed robot image classification model shows the proposed image classification network. The experiments are conducted in section Experiments and analysis. A conclusion is concluded in section Conclusion.

Related works

Image classification is widely used in video surveillance analysis, medical image recognition, face image recognition and other fields. Traditional image classification uses artificial design to extract features, but its generalization ability is poor. Deep learning has been successfully used in speech recognition, natural language processing, and especially computer vision. Deep convolutional neural Networks (DCNN) has become a major research method in the field of computer vision. The following introduces two models including MobileNetV2 network and attention mechanism for this paper.

MobileNetV2 network

With the rapid development of convolutional neural network, its structure is complex, its volume is large, and the calculation amount is large. The demand for hardware resources is bigger. Training cannot be deployed on resource-constrained platforms, so the current research direction is lightweight and high-speed (Zhu et al., 2019). One method is to compress the trained network and get a relatively small model. Another method is to design a small model for training, and MobileNet 2 network is a representative lightweight network.

MobileNetV2 introduces the reverse residual structure and linear bottleneck structure in the network. The reverse residual structure is different from the traditional residual structure. The traditional residual structure reduces the dimension first and then raises it. The reverse residual structure adopts the opposite order. First, 1 × 1 point-by-point convolution of one layer is used to enhance the dimension, and then 3 × 3 deep convolution is used to replace 3 × 3 standard convolution, which can greatly reduce the amount of computation and improve the effectiveness of the network model. Linear bottleneck structure is to design 1 × 1 point with point convolution to reduce the dimension and add it to the input, remove the activation function ReLU in the last layer and replace it with linear activation function. This method can solve the problem of serious information loss. In addition, expansion coefficients are introduced to control the size of the network. The bottleneck structure is shown in Figure 1.

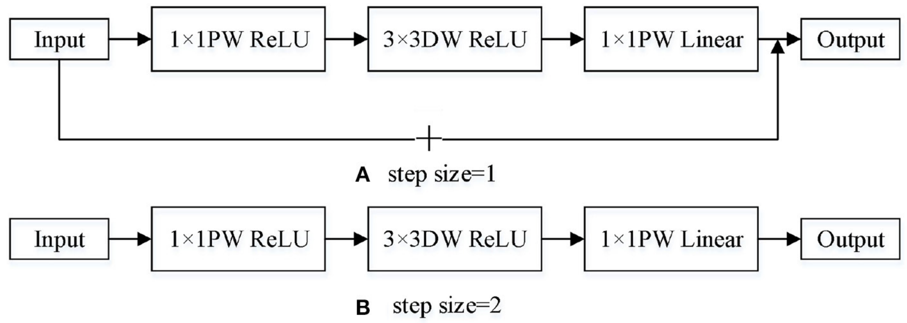

FIGURE 1
MobileNetV2 structure. **(A)** step size=1. **(B)** step size=2.

TABLE 1 Relation between input and output.

Input	Operator	Output
$h \times w \times k$	1 × 1 PW ReLU	$h \times w \times (tk)$
$h \times w \times (tk)$	3 × 3 DW ReLU	$(h/s) \times (w/s) \times (tk)$
$(h/s) \times (w/s) \times (tk)$	1 × 1 PW Linear	$(h/s) \times (w/s) \times (k')$

Table 1 shows the input-output relationship of each layer in the bottleneck diagram in Figure 1. Where k is the number of input channels. h and w are the height and width of the input, respectively. s is the step size. t is expansion coefficient. k' is the number of output channels.

Attention mechanism

When the improvement effect of increasing depth and width in convolutional neural network is not obvious, researchers focus on the attention mechanism. The attention mechanism was first applied to natural language processing. In image classification, the attention mechanism makes the convolutional neural network selectively attach importance to some feature images and suppress some feature images by increasing the weight of some feature images. At present, a lot of researches have been carried out in this area. SENet attention designs a channel module, which introduces two processes of compression and excitation, uses global average pooling to compress the space into 1 × 1 × C feature graphs, and then assigns different weights to different feature maps through two fully connected layers (Shafiq et al., 2020; Li et al., 2022). CBAM focuses on spatial attention on the basis of channel attention, and uses global average pooling and global maximum pooling to reset the weight of channel and spatial attention, which proves to be better than single-channel attention mechanism. In addition, CBAM also explores the effect of the order of attention on the classification effect. Recently, the ECANet module was proposed (Wang Q. L. et al., 2020) based on SENet module, it avoided the reduction of channel dimension and achieved the coverage of local cross-channel interaction by increasing the number of one-dimensional convolution, with certain improvement in classification accuracy.

Proposed robot image classification model

CapsNet neural network

In deep learning, CNN performs very well in image classification and recognition. However, the internal data of CNN does not consider the important spatial hierarchical relationship among features. Although CNN can reduce the size of data space through the network by maximum pooling or adding subsequent convolutional layers, thus increasing the receptive field of upper neurons, that is, obtaining higher-order features in a larger area of the input image, to a certain extent, the model can improve the ability of image recognition, maximum pooling will lose valuable feature information, and traditional convolutional networks cannot solve the fundamental problem of spatial position relationship between low-level features and high-level features. Sabour et al. (2017) proposed CapsNet neural network based on capsule system and dynamic routing mechanism in 2017, which achieved extremely high classification accuracy in MINST data set. Capsules are used to encapsulate and encode the attributes and spatial relations of features in images, such as position, size, direction, deformation, velocity, reflectivity, color and texture, etc. In the multi-layer capsule system, the lower layer capsule sends it to the higher layer capsule that “agrees” with the output result. This step is realized by the dynamic routing algorithm between capsules (Mobiny and Nguyen, 2018).

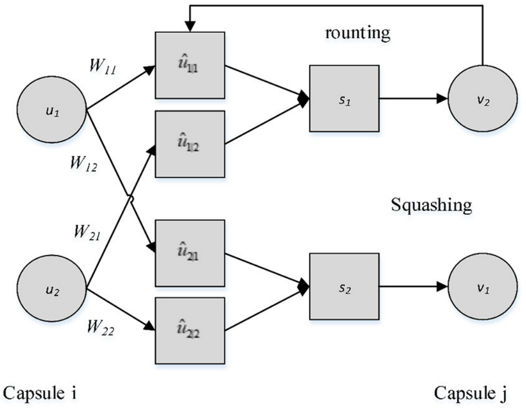

FIGURE 2
The process of transmission between capsules.

Transmission process between capsules

As shown in Figure 2, the input u_i of the CapsNet network model is a vector, which also is a capsule unit. It is similar to the neuron in the convolutional neural network, but the input and output are not scalars but vectors. u_i is multiplied by different weights W_{ij}, respectively, and then it gets the output of the bottom capsule, namely, the prediction vector $\hat{u}_{j|i}$. Then $\hat{u}_{j|i}$ is multiplied by the corresponding coupling coefficient c_{ij}, namely, it conducts weighted sum to get s_j.

$$\hat{u}_{j|i} = W_{ij} u_i \tag{1}$$

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{2}$$

Where, c_{ij} is the coupling coefficient determined by the iterative dynamic routing process, and there is no bias term in the capsule network model.

The non-linear activation function Squashing is used in CapsNet to scale a vector between 0 and 1. After s_j is obtained, the Squashing function is used to transform vector s_j into vector v_j, as shown in Equation (3):

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||} \tag{3}$$

Since this function is non-linear, the output v_j preserves the dimension of s_j and does not change the direction of the vector, only the changes the vector size. The length of its output vector represents the probability of a given feature detected by the capsule. Therefore, the probability of each capsule is between 0 and 1.

Dynamic routing algorithm between capsules

The updating iteration of coupling coefficient c_{ij} in Capsule network model is realized by dynamic routing algorithm. The input of the dynamic routing algorithm is the prediction vector $\hat{u}_{j|i}$ and the number r of routing iterations. b_{ij} is a temporary variable whose value will be updated in the iteration process. When the whole dynamic routing algorithm is finished, its value will be saved to c_{ij}. b_{ij} is initialized to zero at the beginning of the training (Cao et al., 2020; Madhu et al., 2021; Sepas-Moghaddam et al., 2021).

The following process is iterated r times. Firstly, softmax is used to calculate the value of weight c_{ij}, as shown in Equation (4).

$$c_{ij} = \frac{e^{b_{ij}}}{\sum_k e^{b_{ik}}} \tag{4}$$

Softmax can ensure that all the weight c_{ij} is non-negative and the sum is equal to 1.

The weighted sum of $\hat{u}_{j|i}$ is then used to obtain s_j, and the squashing function is used to obtain v_j. The last step of the iteration is to check each high-level capsule input and update the corresponding weight b_{ij}.

$$b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} \cdot v_j \tag{5}$$

After updating the weights, the algorithm returns and restarts the calculation of c_{ij}, and repeating r times. Dynamic routing algorithm is easy to converge, but it also has the problem of over-fitting. Although increasing the number of iterations can improve the accuracy, it will increase the generalization error, so it is not suitable for too many iterations.

Two methods are used to optimize and improve dynamic routing. One is to introduce Adam function, the other is to improve the compression function. Shafiq et al. (2020) added Adam algorithm in the dynamic routing, which could make the model convergence more stable. The algorithm averages the moving index of the parameter error between capsules, dynamically adjusts the Leaning Rate of the routing algorithm, and adjusts the weight b_{ij}.

In this paper, Adam algorithm is optimized, and dynamic routing is improved. During iteration, Sigmoid after translation is multiplied with $s_j/||s_j||$ to replace Squashing function, and Squashing function is used in the final output. The improved algorithm is as follows as shown in Algorithm 1:

For each category k appearing in the image, the capsule uses a separate profit loss L_k, as shown in Equation (6).

$$L_k = T_k \max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k) \max(0, ||v_k|| - m^-)^2 \tag{9}$$

If there are k images, it sets $T_k = 1$, and $m^+ = 0.9$, $m^- = 0.1$. The value of λ is 0.5 to reduce the loss when some categories do

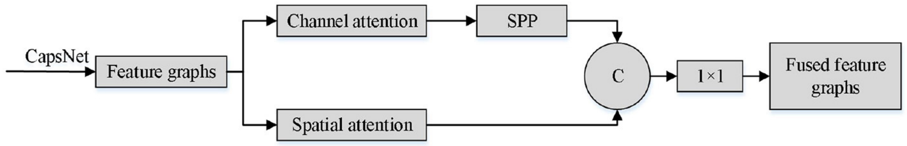

FIGURE 3
Proposed C-MobileNetV2 structure.

```
Procedure $I - ROUTING(\hat{u}_{j|i}, r, l)$
for all capsule $i$ in layer $l$ and $j$ layer $(l+1)$
for r iterations do
    $c_i \leftarrow soft\max(b_i)$    (6)
    $\widehat{m}_t \leftarrow m_t/(1-\alpha^t)$    (7)
    $\widehat{n}_t \leftarrow n_t/(1-\beta^t)$    (8)
Output: return squash $(v_j)$
```

Algorithm 1. Modified dynamic routing.

not appear, so as to ensure that the vector modulus representing the digital capsules of this category is as large as possible, and the vector modulus of other categories is as small as possible to achieve accurate classification. The total loss is the sum of all the digital capsule losses.

The definition of softmax function is given by formula (4), the definition of Squashing function is given by formula (3), and the Sigmoid function is as follows:

$$sigmoid(||s_j||) = 1/(1 + e^{-||s_j||}) \tag{10}$$

Because CapsNet network finally measures the probability of a result by the modulus size of the vector. The magnitude of the modulus is directly related to the probability. The larger modulus denotes the greater corresponding probability. The Sigmoid function converges faster than the Squashing function. Therefore, using the translation Sigmoid function multiplied with $s_j/||s_j||$ to replace the Squashing function in the iterative link can play an amplification effect when the result is close to 0 or 1. So it can improve the recognition accuracy of some specific kinds of images, and improve the efficiency of model operation.

CapsNet based on MobileNetV2 network

In view of the requirements of image classification for the lightweight and accuracy of deep convolutional neural network, this paper takes the lightweight network MobileNetV2 as the benchmark, and then outputs the feature map from CapsNet into MobileNetV2. On the basis of ensuring the lightweight, the self-designed attention mechanism structure is added to improve the classification accuracy of the network. In the process of channel attention and spatial attention, small convolution blocks are used to ensure the computational requirements of the network. Furthermore, convolutional blocks of different sizes are designed to extract multi-scale information of feature graphs, and convolutional blocks with different dilated rates are designed to make the network pay attention to global features. Finally, it adds the designed attention module to the MobileNetV2 network structure.

A new image classification structure C-MobileNetV2 is proposed in this paper, and its structure is shown in Figure 3. First, it inputs the feature graphs extracted by CapsNet into MobileNetV2. Feature graphs enter the channel and spacial modules in parallel. Then, the multi-scale semantic information is extracted through the spatial pyramid pooling (SPP) layer (Tai et al., 2020; Wang Q. L. et al., 2020) in the channel attention module and enters into the multi-layer perceptron. Under the action of activation function, the feature graph after redistributing weight is obtained. In the spatial attention module, the convolution of different scales is designed to increase the receptive field, and the spatial feature map with comprehensive semantic and distinct characteristics is obtained. Finally, the feature graphs generated by the two modules are fused to generate the extracted feature graphs.

Because MobileNetV2 introduces the reverse residual structure and linear bottleneck structure, the computational cost is greatly reduced. This paper introduces self-designed attention module to improve its classification accuracy. The improved network structure is shown in Table 2. The designed attention module is integrated into the third, fifth, sixth, and seventh layers of the network structure. The purpose of the design is to increase the recognition accuracy without changing the size of the network and the amount of computation.

TABLE 2 Parameters in C-MobileNetV2 structure.

Layer	Input	Operator	t	c	n	s	C-MobileNetV2
Input layer	224 × 224 × 3	conv2d	–	64	1	2	×
1	112 × 112 × 34	Bottleneck	1	16	1	1	×
2	112 × 112 × 16	Bottleneck	6	24	2	2	×
3	56 × 56 × 24	Bottleneck	6	32	3	2	√
4	28 × 28 × 32	Bottleneck	6	64	4	2	×
5	14 × 14 × 64	Bottleneck	6	96	3	1	√
6	14 × 14 × 96	Bottleneck	6	160	3	2	√
7	7 × 7 × 160	Bottleneck	6	320	1	1	√
8	7 × 7 × 320	conv2d 1 × 1	–	1,280	1	1	–
9	7 × 7 × 1,280	Avgpool 7 × 7	–	–	1	–	–
Output layer	1 × 1 × 1,280	conv2d 1 × 1	–	100	–	–	–

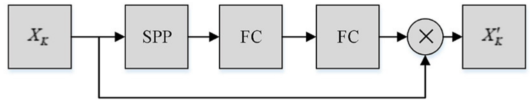

FIGURE 4
Channel attention.

Channel attention module

Channel attention module is shown in Figure 4. When the input feature graph $X_K \in R^{H \times W \times C}$ is entered, Through spatial pyramid pooling, global information is compressed by H × W in spatial dimension. The channel feature graph is compressed into 1 × 1 × C, and the corresponding weight information is generated by mapping. The spatial pyramid adaptively and evenly stores the input feature graphs into three scales. The output of the three channels is adjusted to three one-dimensional vectors, and the one-dimensional attention graph is generated after fusion.

In order to assign different weights to different channels, a one-dimensional attention diagram is extracted from the pooling layer of spacial pyramid. The two fully connected layers are followed by the sigmoid function to normalize the output to the range of (0,1). The output after SPP pooling layer and two full connection layers is:

$$\tilde{M} = sig(W_2 \rho(W_1 M)) \tag{11}$$

Where $\tilde{M}$ is the output after SPP pooling layer and two fully connection layers. W_1 and W_2 are the fully connected layer 1 and layer 2, respectively. M represents the one-dimensional feature graph generated after passing through the SPP layer. ρ denotes ReLU.

The output of spatial channel attention mechanism is:

$$X_K = X_K \otimes \tilde{M} \tag{12}$$

SPP is essentially multiple average pooling layer. For feature graphs $a \times a$ with different sizes, the output $n \times n$ with fixed size can be generated by adjusting the size and step of slide window. For example, Figure 5 is composed of three average pooling layers. For the input with any size, feature vectors of 4 × 4, 2 × 2 and 1 × 1 are extracted. The feature vectors of 21 dimensions are uniformly output through feature fusion. So that the size of the input image is not limited and the multi-scale semantic features can be extracted.

Spatial attention module

In order to highlight significant spatial features, suppress unimportant features and improve the classification accuracy of the network on the basis of increasing the receptive field, parallel input of spatial information and channel information is designed in this paper. The spatial attention module is composed of two 3 × 3 dilated convolution layers and two 1 × 1 convolution layers.

The spatial attention module is shown in Figure 6. Firstly, the feature graph is compressed and it passes through two dilated convolution layers with dilated rates 2 and 3, respectively. Finally, the one-dimensional feature graph is generated after compression. The reason for choosing two dilated convolution layers is to maximize the receptive field. However, if only the convolution blocks with large dilated rate are selected, the feature semantics will be spatially discontinuous and the extracted features will be too scattered, which will lead to a decrease rather than an increase in terms of the classification accuracy. Therefore, two 3 × 3 dilated convolution blocks with dilated rates 2 and 3 are selected, so that the two convolution blocks can increase the size of the receptive field and extract continuous feature semantics after superposition. The actual

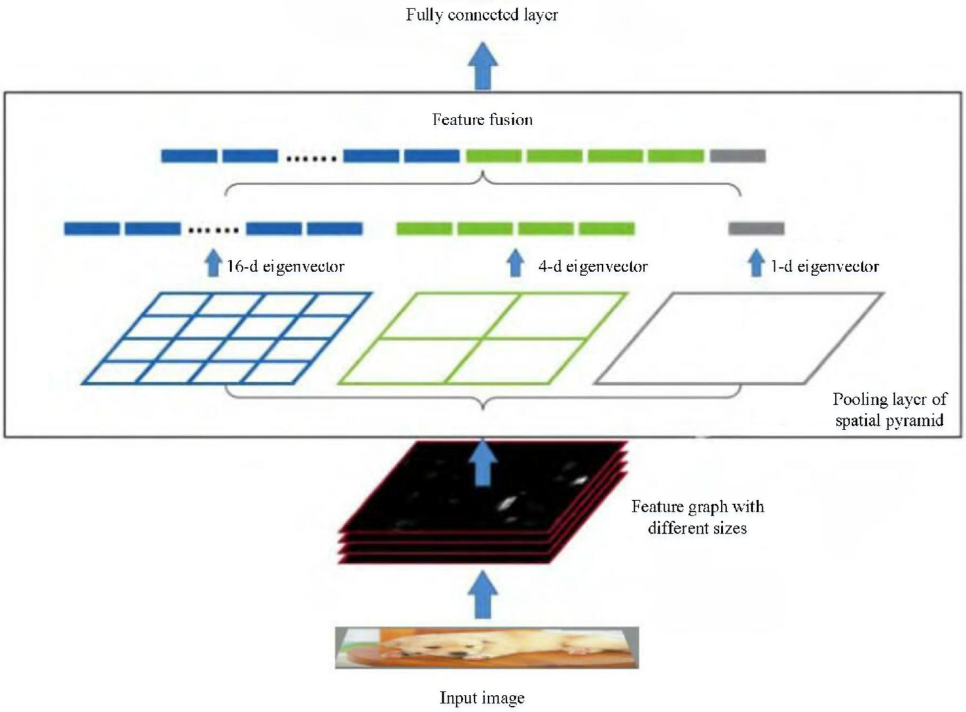

FIGURE 5
SPP structure.

receptive field size of the two dilated convolution layers is:

$$F_{i+1} = F_i + (k-1) \times S_i \tag{13}$$

Where, F_{i+1} is the receptive field of the current layer. F_i is the receptive field of the upper layer. S_i is the product of step sizes of all previous layers (excluding current layer). K is the convolution kernel size. After calculation, the actual receptive field size is 12 × 12, represented by $f^{12\times 12}$. The output of spatial attention module is:

$$X'_K = X_K \otimes \tilde{V} \tag{14}$$

Where $\tilde{V}$ represents the attention graph generated after dilated convolution, and its calculation is as follows:

$$\tilde{V} = sig(f^{12\times 12}(A_{avg}))_{avg} \tag{15}$$

Where A_{avg} is the output of the input feature graph after 1 × 1 convolution. sig is the sigmoid function.

In our proposed CapsNet model, there is a convolution layer with a convolution kernel size of 7 × 7 and step size of 2, which has ReLU non-linear activation function. The second layer is also the convolution layer. The convolution kernel size is 6 × 6 and the step size is 1. Then, a pooling layer is connected, and the pooling core size is 2 × 2, step size is 1, and maximum pooling is adopted. The pooling layer is then connected to a 6 × 6 with step size 1, and the data is reconstructed and then input to the PrimaryCaps layer, so that the main feature information of the image can be obtained as much as possible. Due to the large amount of data, pooling layer is introduced to reduce data dimension to improve model performance and running efficiency.

Experiments and analysis

The operating system of this algorithm is Windows 10 and GPU is TeslaT4 (64 GB video memory). The deep learning framework adopts PyTorch, and the verification environment is PyCharm+Anaconda.

The experiment data set is cifar-100 benchmark data set. This data set is widely used in the validity of classification algorithms. There are a total of 60,000 color images with size 32 × 32 in the dataset, they are divided into 100 categories, each category contains 600 images. These 100 classes belong to 20 superclasses. In the dataset, 50,000 images are used for training and 10,000 images for testing. The testing parameters are set as follows. The gradient descent method is selected as the optimization algorithm. The batch size is 512. The initial learning rate is 0.1. Training batch size is 100. Momentum is 0.9. The number of iterations is 300. The weight decay is 5×10^{-4}.

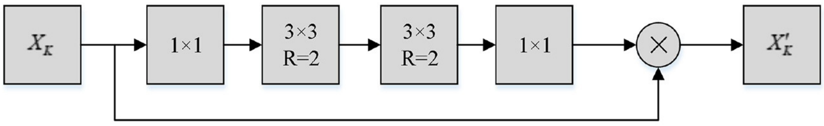

FIGURE 6
Spatial attention.

TABLE 3 Comparison with different methods.

Method	Classification accuracy/%	Error/%	Running time/s
MobileNetV2	71.69	1.31	0.55
BAVCT	79.84	0.41	0.88
TRk-CNN	86.71	0.23	0.72
Feat-WCLTP	90.37	0.19	0.61
C-MobileNetV2	96.58	0.02	0.58

TABLE 4 Comparison of classification accuracy of different modules/%.

Module	+SENet	+CBAM	+ECANet	+Capsule
VGG16	73.24	73.72	73.82	74.24
ResNet18	76.03	76.14	76.30	76.46
DenseNet	75.59	75.90	75.93	76.31
MobileNetV2	75.88	75.96	76.02	76.65

We select four state-of-the-art robot image classification methods [BAVCT (Liu et al., 2022), TRk-CNN (Jun et al., 2021), Feat-WCLTP (Shamaileh et al., 2020)] and the traditional MobileNetV2 to make comparison with proposed method (C-MobileNetV2). The classification effect and error results are shown in Table 3.

The classification accuracy with C-MobileNetV2 is 96.58%, which is improved by 16.74, 9.87, and 6.21% than that of BAVCT, TRk-CNN, Feat-WCLTP, respectively. Also the error is the lowest. The C-MobileNetV2 can achieve better image classification effect. The running time is 0.58 s with C-MobileNetV2, which is higher that MobileNetV2 with 0.55 s due to the introduction of capsule module in MobileNetV2. However, it is lower than that of BAVCT (0.88 s), TRk-CNN (0.72 s), Feat-WCLTP (0.61 s).

In order to verify the designed attention network performance, MobileNetV2 based on capsule network is compared with different classical convolutional neural networks. Four typical neural networks are selected including VGG16 network, ResNet18 residual network, DenseNet networks. The classification effect and improvement effect are shown in Table 4.

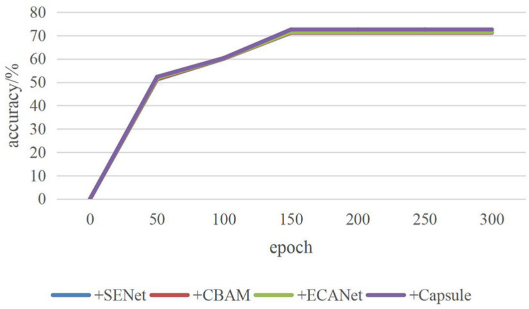

FIGURE 7
Accuracy change curve.

As can be seen from Table 4, for different convolutional neural networks, the classification effect of CBAM module with dual attention mechanism is better than that of SENet module with single attention. The classification accuracy of the newly proposed ECANet module is higher than that of CBAM module, and the C-MobileNetV2 in this paper is better than that of ECANet module.

In the process of 300 iterations, the accuracy changes of the four attention modules in VGG16 are shown in Figure 7. The solid yellow line represents the accuracy changes of C-MobileNetV2 module in each convolutional neural network. It can be seen that the solid yellow line is superior to other colored solid lines in the iterative process. It is similar to ResNet18, DenseNet and MobileNetV2.

In this paper, cross-validation is also used to illustrate the effectiveness of adding attention mechanism module, and the experiment results are shown in Table 5. In here, CAM: channel attention module, SAM: spacial attention module.

Experiment results show that although the traditional MobileNetV2 network has a poor classification effect on the complex data set CIFAR-10, the capsule network with the attention mechanism can improve the accuracy of the complex data set.

The above data analysis effectively proves the effectiveness and superiority of C-MobilenetV2 network, especially the classification accuracy has been greatly improved. The main reasons include two aspects. First, channel attention breaks the shackles of global average pooling and global maximum pooling. It introduces SPP into module, which makes feature extraction

TABLE 5 Classification error rates with different improved modules/%.

MobileNetV2	7.11
Capsule	5.54
MobileNetV2+CAM+SAM	5.52
Capsule+CAM+SAM	4.63
MobileNetV2+Capsule (CAM+SAM)	**3.78**

They denote the best values obtained by proposed method.

TABLE 6 Experimental results of different models on MORPH Album2.

Model	MAE
MobilenetV2	3.52
C-MobilenetV2	**3.05**

They denote the best values obtained by proposed method.

TABLE 7 Experimental results of different models on Adience/%.

Model	Average accuracy	1-off
MobilenetV2	68.67	93.43
C-MobilenetV2	**72.98**	**93.85**

They denote the best values obtained by proposed method.

more hierarchical, weight distribution more reasonable, and feature characterization ability optimization obvious. Second, the dilated convolution is designed in spatial attention, and the pooling block size and dilated rate size are selected reasonably. Compared with the traditional spatial attention module, C-MobilenetV2 can increase the receptive field, further expand the extraction range and take into account the problem of semantic continuity, achieving a good dynamic balance and extracting saliency feature images more reasonable. The superposition of the optimization effects of the two aspects improves the classification accuracy of the proposed network in this paper significantly.

We also conducts experiments and analysis on high-resolution face age datasets. Face image age estimation has become a very important task in the field of pattern recognition and computer vision, which has a wide range of practical value. Face age estimation is more difficult than the general image classification task because of the small inter-class difference between adjacent ages.

In order to verify the effectiveness of proposed network in high-resolution and challenging image data sets, this paper uses MobilenetV2 as the basic network to conduct experiments on MORPH Album2 and Adience data sets. The experimental results are shown in Tables 6, 7. As can be seen from Table 6, in the MORPH Album2 data set, the MAE value of C-Mobile NETv2 in this paper is significantly reduced. As can be seen from Table 7, in the age group classification comparison experiment conducted on Adience, an unrestricted age data set, the age estimation accuracy obtained by C-MobilenetV2 is higher than the classification accuracy of MobilenetV2 network at the same level. The experimental results show that C-MobilenetV2 network has better learning ability than the original MobilenetV2 network on MORPH Album2 and Adience high resolution data sets, which verifies the adaptability of C-MobilenetV2 in different types and different resolutions of data sets.

Conclusion

The MobileNetV2 combined with capsule network model can effectively solve the problems of insufficient feature extraction of original MobileNetV2 network and poor performance in complex data sets. In the feature extraction process, we first adopt capsule network to obtain feature graph. Based on the lightweight MobileNetV2 network, a channel-space dual collaborative attention module is designed, which is embedded in MobileNetV2 network structure and successfully applied to image classification. On the CIFAR-100 public standard data set, different classical convolutional neural networks and different attention modules are embedded for comparison, which achieves good results and greatly improves the classification accuracy, it can fully confirm the effectiveness of the proposed algorithm. Future works will be researched in the area of image classification with advanced deep learning methods.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JZ implemented the code and draft the manuscript. XY and XL assisted to implement the code and discussed the manuscript. CW guided the research and discussed the results. JZ guided the research, implemented parts of code, and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. U20A20197, 61973063, U1713216, 61901098, and 61971118, Liaoning Key Research and Development Project 2020JH2/10100040, the China Postdoctoral Science Foundation 2020M670778, the

Northeastern University Postdoctoral Research Fund 20200308, the Scientific Research Foundation of Liaoning Provincial Education Department LT2020002, the Foundation of National Key Laboratory (OEIP-O-202005), and the Fundamental Research Fund for the Central Universities of China N2026005, N181602014, N2026004, N2026006, N2026001, and N2011001.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akay, M., Du, Y., Sershen, C. L., Wu, M. H., Chen, T. Y., Assassi, S., et al. (2021). Deep learning classification of systemic sclerosis skin using the MobileNetV2 model. *IEEE Open J. Eng. Med. Biol.* 2, 104–110. doi: 10.1109/OJEMB.2021.3066097

Branch, M., and Carvalho, A. S. (2021). Polyp segmentation in colonoscopy images using U-Net-MobileNetV2. *arXiv*:2103.15715. doi: 10.48550/arXiv.2103.15715

Cao, J., Zhang, J., and Huang, W. (2021). Traffic sign detection and recognition using multi-scale fusion and prime sample attention. *IEEE Access* 9, 3579–3591. doi: 10.1109/ACCESS.2020.3047414

Cao, S., Yao, Y., and An, G. (2020). E2-capsule neural networks for facial expression recognition using AU-aware attention. *IET Image Process.* 14, 2417–2424. doi: 10.1049/iet-ipr.2020.0063

Choi, K., Oh, S., and Sohn, C. (2020). Application of OpenPose and deep learning for intelligent surveillance reconnaissance system. *J. Adv. Military Stud.* 3, 113–132. doi: 10.37944/jams.v3i3.80

He, K. M., Zhang, X. Y., Ren, S. Q., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. H. (2020). Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2011–2023. doi: 10.1109/TPAMI.2019.2913372

Hui, T.-W., Tang X, and Loy, C. C. (2018). "LiteFlowNet: a lightweight convolutional neural network for optical flow estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 8981–8989.

Jisi, A., and Yin, S. (2021). A new feature fusion network for student behavior recognition in education. *J. Appl. Sci. Eng.* 24, 133–140. doi: 10.6180/jase.202104_24(2).0002

Jun, T. J., Eom, Y., Kim, D., and Kim, C. (2021). TRk-CNN: transferable ranking-CNN for image classification of glaucoma, glaucoma suspect, and normal eyes. *Expert Syst. Appl.* 182, 115211. doi: 10.1016/j.eswa.2021.115211

Li, P., Laghari, A. A., Rashid, M., Gao, J., Gadekallu, T. R., Javed, A. R., et al. (2022). "A deep multimodal adversarial cycle-consistent network for smart enterprise system," in *IEEE Transactions on Industrial Informatics*.

Liu, Y., Dou, Y., Jin, R. C., Li, R. C., and Qiao, P. (2022). Hierarchical learning with backtracking algorithm based on the Visual Confusion Label Tree for large-scale image classification. *Visual Comput.* 38, 897–917. doi: 10.1007/s00371-021-02058-w

Madhu, G., Govardhan, A., Srinivas, B. S., Sahoo, K. S., Jhanjhi, N. Z., Vardhan, K. S., et al. (2021). Imperative dynamic routing between capsules network for malaria classification. *Comput. Mater. Continua* 680, 903–919. doi: 10.32604/cmc.2021.016114

Mobiny, A., and Van Nguyen, H. (2018). "Fast capsnet for lung cancer screening," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science(), Vol 11071*, eds A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger (Cham: Springer), 741–749. doi: 10.1007/978-3-030-00934-2_82

Prabhu, N. L., Jun, D. L. J., Dananjaya, P. A., Wen, S. L., and Raghavan, N. (2020). Exploring the impact of variability in resistance distributions of RRAM on the prediction accuracy of deep learning neural networks. *Electronics* 9, 414. doi: 10.3390/electronics9030414

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Adv. Neural Information Process. Syst.* 3857–3867. Available online at: https://proceedings.neurips.cc/paper/2017/hash/2cad8fa47bbef282badbb8de5374b894-Abstract.html

Sandler, M., Howard, A., Zhu, M. L., Zhmoginov, A., and Chen, L. C. (2018). "MobileNetV2: inverted residuals and linear bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT).

Sepas-Moghaddam, A., Etemad, A., Pereira, F., and Correia, P. L. (2021). CapsField: light field-based face and expression recognition in the wild using capsule routing. *IEEE Trans. Image Process.* 30, 2627–2642. doi: 10.1109/TIP.2021.3054476

Shafiq, M., Tian, Z., Bashir, A. K., Du, X., and Guizani, M. (2021). CorrAUC: a malicious Bot-IoT traffic detection method in IoT network using machine-learning techniques. *IEEE Internet Things J.* 8, 3242–3254. doi: 10.1109/JIOT.2020.3002255

Shafiq, M., Tian, Z., Bashir, A. K., and Jolfaei, A. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: a survey. *Sustain. Cities Soc.* 60, 102177. doi: 10.1016/j.scs.2020.102177

Shamaileh, A. M., Rassem, T. H., Chuin, L. S., and Sayaydeh, O. N. A. (2020). A new feature-based wavelet completed local ternary pattern (Feat-WCLTP) for texture image classification. *IEEE Access* 8, 1. doi: 10.1109/ACCESS.2020.2972151

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*:1409.1556. doi: 10.48550/arXiv.1409.1556

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA), 1–9.

Tai, S. K., Dewi, C., Chen, R. C., Liu, Y. T., Jiang, X. Y., and Yu, H. (2020). Deep learning for traffic sign recognition based on spatial pyramid pooling with scale analysis. *Appl. Sci.* 10, 6997. doi: 10.3390/app10196997

Wang, Q. L., Wu, B. G., Zhu, P. F., Li, P. H., Zuo, W. M., and Hu, Q. H. (2020). "ECA-Net: efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA), 11531–11539.

Wang, Q. L., Yin, S. L., Sun, K., Li, H., Liu, J., and Karim, S. (2020). GKFC-CNN: modified Gaussian Kernel Fuzzy C-means and convolutional neural network for apple segmentation and recognition. *J. Appl. Sci. Eng.* 23, 555–561. doi: 10.1109/CVPR42600.2020.01155

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). "CBAM: convolutional block attention module," in *European Conference on Computer Vision* (Munich), 3–19. doi: 10.1007/978-3-030-01234-2_1

Yang, B., Liu, Z., Xing, Y., and Luo, C. (2011). "Remote sensing image classification based on improved BP neural network," *2011 International Symposium on Image and Data Fusion* (Tengchong), 1–4.

Yin, S., and Li, H. (2020). Hot region selection based on selective search and modified fuzzy C-means in remote sensing images. *IEEE J. Selected Top. Appl. Earth Observ. Remote Sens.* 13, 5862–5871. doi: 10.1109/JSTARS.2020.3025582

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* 100, 104190. doi: 10.1016/j.engappai.2021.104190

Zhu, J., Chen, T., and Cao, J. (2019). Siamese network using adaptive background superposition initialization for real-time object tracking. *IEEE Access* 7, 119454–119464. doi: 10.1109/ACCESS.2019.2937166