



OPEN

A study on expression recognition based on improved mobilenetV2 network

Qiming Zhu^{1,4}, Hongwei Zhuang^{1,4}✉, Mi Zhao², Shuangchao Xu¹ & Rui Meng³

This paper proposes an improved strategy for the MobileNetV2 neural network (I-MobileNetV2) in response to problems such as large parameter quantities in existing deep convolutional neural networks and the shortcomings of the lightweight neural network MobileNetV2 such as easy loss of feature information, poor real-time performance, and low accuracy rate in facial emotion recognition tasks. The network inherits the characteristics of MobileNetV2 depthwise separated convolution, signifying a reduction in computational load while maintaining a lightweight profile. It utilizes a reverse fusion mechanism to retain negative features, which makes the information less likely to be lost. The SELU activation function is used to replace the RELU6 activation function to avoid gradient vanishing. Meanwhile, to improve the feature recognition capability, the channel attention mechanism (Squeeze-and-Excitation Networks (SE-Net)) is integrated into the MobileNetV2 network. Experiments conducted on the facial expression datasets FER2013 and CK + showed that the proposed network model achieved facial expression recognition accuracies of 68.62% and 95.96%, improving upon the MobileNetV2 model by 0.72% and 6.14% respectively, and the parameter count decreased by 83.8%. These results empirically verify the effectiveness of the improvements made to the network model.

Keywords Expression recognition, MobileNetV2, Reverse fusion, Attention mechanism, SELU

Facial expression recognition (FER)¹ is one of the key research directions in computer vision. It has been found that facial expressions account for more than 55% of the emotional information conveyed by humans², and in populations with relatively poor language skills such as newborns and the elderly, expressions contain even richer information. Currently, FER technology has been widely used in numerous sectors including human–computer interaction³, pain identification⁴, fatigue driving judgment⁵, and criminal interrogation⁶.

Facial expression recognition generally includes three steps: preprocessing⁷, feature extraction⁸, and expression classification⁹. Traditional methods for feature extraction implemented designs such as local binary patterns¹⁰, directional gradient histograms¹¹, and scale-invariant feature transformation¹². However, the above methods are complex in design and feature extraction information is incomplete. At present, researchers have designed a variety of deep learning network models for application in multiple domains such as image processing¹³, natural language processing¹⁴, and speech recognition¹⁵, and have achieved good classification results in face expression recognition. Typical models include VGG¹⁶, RESNET¹⁷, ALEXNET¹⁸, GoogLeNet¹⁹, MobileNetV1²⁰, MobileNetV2²⁰, MobileNetV3²¹ and other models. In MobileNetV1, the size and computational complexity of the network structure was reduced by replacing the standard convolution with a Depthwise Separable Convolution²². On the other hand, MobileNetV2 adds multiple pointwise convolutions while maintaining the depthwise-separable convolutions to further improve the performance of the network structure and reduce the computational complexity. However, when the network structure of MobileNet series is applied to the task of image classification, due to the internal design of the network, optimization strategy and activation function, it will constantly ignore the negative feature information in the input information, which may be the key point of image classification. So, Improvement methods for MobileNet network structure in the field of image classification have been successively proposed in recent years. For example, in 2019, Yonis Gulzar²³ designed a specific five-layer in mobilenetv2 network while retaining the pre-trained model using migration learning, and achieved good results in automatically extracting fruit features for recognition. Yue Pang²⁴ designed an

¹College of Equipment Support and Management, Engineering University of PAP, Xi'an 710086, China. ²Basic Education, Engineering University of PAP, Xi'an 710086, China. ³College of Military Basic Education, Engineering University of PAP, Xi'an 710086, China. ⁴These authors contributed equally: Qiming Zhu and Hongwei Zhuang. ✉email: zhuanghw01@163.com

improved mobilenetv2 network in order to solve the problem of sheep recognition and tracking in large-scale sheep farming, while a series of validation tests were carried out. The algorithm utilizes facial features to recognize individual sheep, and the model has the highest accuracy rate among similar algorithms. B. Anil Kumar²⁵ et al. added five different layers to the pre-trained MobileNetV2 architecture in order to obtain better classification accuracy with fewer training parameters for a given face detection data. The experimental results show that the method achieves 99.64% recognition accuracy for photo images and higher accuracy for real time video images. Literature²⁶ introduced the attention module into the MobileNetV1 model to enhance local feature extraction of facial expressions. Then, the center loss and cross-entropy loss are combined to optimize the model parameters to reduce the intra-class distance and increase the inter-class distance, and the experimental results show the effectiveness of the method.

This paper proposes the following improvements based on previous studies.

- (i) Channel attention mechanism with residual network structure is used in the feature extraction stage to extract features from the image.
- (ii) In order to retain the negative value information in the features, inverse and fusion operations are performed on the feature map, which is conducted to the next layer by deep fusion convolution to avoid the loss of negative value information when passing through the activation function.
- (iii) SELU activation function is used instead of RELU6 activation function for the operation to improve the feature extraction ability of the model.

Database and model

Dataset selection

The publicly available datasets FER2013 and CK+ were used as the datasets used for training and testing. FER2013²⁷ Expression Dataset Created by Carrier and Courville for the ICML2013 Facial Expression Recognition Competition. The FER2013 dataset contains 35,886 face expressions, including face expressions of different age groups in daily life, the pixel values of the images are all 48×48 , and they are categorized into seven types of expressions: angry, disgust, fear, happy, sad, surprise and neutral. It is divided into three sections: training, validation and testing. Training section contains 28,709 images, validation section contains 3589 images and testing section contains 3589 images. Each image was labeled with the correct facial expression category. The highest human recognition accuracy on the FER2013 dataset was only 65% to 70% due to acquisition and labeling errors. The CK+²⁸ database is an extension of the Cohn-Kanade Dataset, which consists of seven expressions: anger, tempt, disgust, fear, happy, sadness, and surprise, and contains 123 participants and 593 image sequences. The CK+ dataset is a more general face expression dataset, which is suitable for the research of face expression recognition. Figure 1 represents some samples from the FER dataset and the CK+ expression dataset.

Deepwise separable convolution

AlexNet won first place in the 2012 ImageNet competition using Convolutional Neural Networks (CNN) for image classification. Subsequent researchers have designed many more excellent network models, such as VGGNet16/19, GoogLeNet, ResNet50, etc., which have achieved good advantages over traditional classifiers. However, with the deepening of the network, the model computation requires more and more resources, often need to operate on specific machines, which limits the use of deep learning models. Therefore, in 2017, Google Inc. proposed the MobilenetV1 lightweight network model, MobilenetV1 makes full use of the computational resources while maximizing the accuracy of the model in 2018, Google Inc. proposed the mobilenetV2 model based on MobilenetV1.

The traditional standard convolutional operations are characterized as shown in Fig. 2, Suppose the input feature map size is $M \times M$, the number of channels is C , the convolution kernel size of standard convolution is $N \times N$, the number of K . And assuming that the output is the same size as the input, the output size after the standard convolutional kernel is $M \times M$, and the number of output channels is K . Traditional standard convolution actually consists of two steps: filtering the feature map first, combining the filtered results, Fig. 1b shows the standard convolution of the input feature map with the $i(1 \leq i \leq k)$ convolution kernel. In this process, each channel in the input feature map is first convolved with each channel of the corresponding convolution kernel, The result of the convolution forms a single-channel feature map of C of $M \times M$, and then this C and the result are merged to form a final single feature map of $M \times M \times 1$. Since there are K convolution kernels, the input



Figure 1. Example of FER2013²⁸ and CK+²⁹ images.

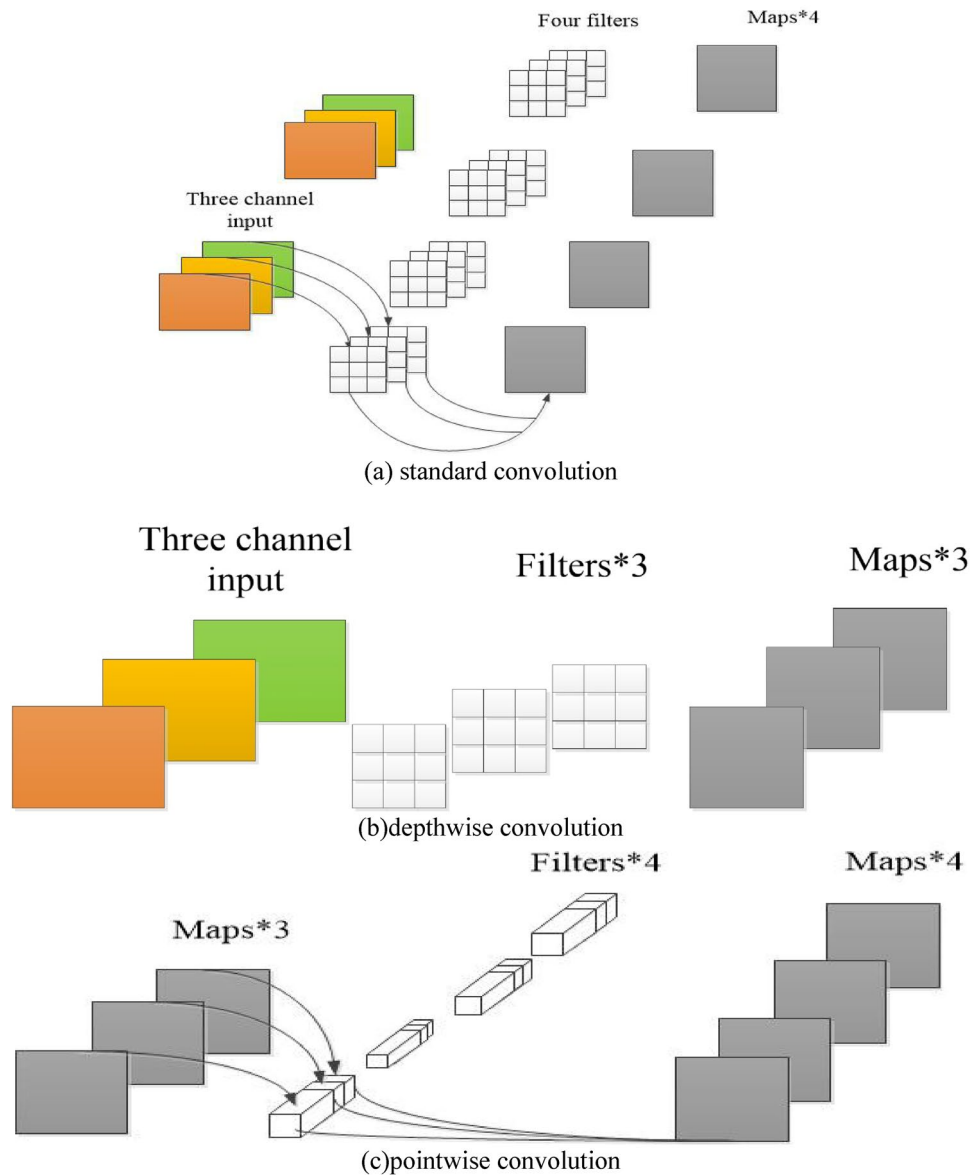


Figure 2. Schematic of the three convolutions.

feature map has a total of K of $M \times M \times 1$ results after standard convolution with all K convolution kernels, The final result is the output feature map of $M \times M \times K$, as shown in Fig. 2a. The standard convolution is computed:

$$M \times M \times N \times N \times C \times K \tag{1}$$

Depthwise separable convolution decomposes the standard convolution into a depth convolution kernel a pointwise convolution, the depthwise convolution process actually convolves each channel of the input with its corresponding convolution kernel individually, and finally the resulting convolution result corresponding to each channel is used as the final depthwise convolution result. In fact, the process of depthwise convolution completes the filtering of the input feature map, and the depthwise convolution process is shown in Fig. 2b, whose computation is:

$$M \times M \times N \times N \times C \tag{2}$$

Here, the pointwise convolution takes the result of depth convolution as input, the convolution kernel size is 1×1 , and the number of channels is the same as the input. The process of point convolution is similar to standard convolution, which is actually a linear combination of each pixel point on different channels, and retains the original planar structure of the image, regulating the depth. Compared with depth convolution, dot convolution has the ability to change the number of channels, which can accomplish the function of dimension upgrading or downgrading. Dot convolution is shown in Fig. 2c, and its computation is:

$$M \times M \times 1 \times 1 \times C \times K = M \times M \times C \times K \tag{3}$$

The total computation of the deepwise separable convolution is:

$$M \times M \times N \times N \times C + M \times M \times C \times K \tag{4}$$

Deepwise separable convolutions versus traditional standard convolutional computations:

$$\frac{M \times M \times N \times N \times C + M \times M \times C \times K}{M \times M \times N \times C \times K} = \frac{1}{K} + \frac{1}{N^2} \tag{5}$$

From the above equation, it can be seen that the depth separable convolution can effectively reduce the amount of computation, if the network uses 3*3 convolution kernel size for convolution, the depth separable convolution can reduce the amount of computation by 8 to 9 times. Compared with the traditional convolutional computation, this approach is less computationally intensive and can effectively extract features with less loss of accuracy.

The researchers introduced a width factor α into the Mobilenet family of networks. α has the effect of regulating the input and output channels so that the input channel is changed from M to αM and the output channel is changed from N to αN . Compared with the standard convolutional layer, the number of parameters is reduced by α^2 . The value of α is taken in the range of (0,1], and it is usually set to 1.0, 0.75, 0.5 or 0.25.

Reverse fusion methods(RFM)

In this paper, we design a deep convolutional inverse layer and a deep convolutional fusion layer to solve the problems of insufficient structural feature extraction and easy to ignore negative feature information in MobileNetV2 network. The original image features and their inverse features are fused and extracted, the class residual structure is added, the convolutional blocks are combined, and the whole network is optimized by loss function for classification. Figure 3 shows a schematic diagram of the reverse fusion operation. The feature map obtained from the deep convolutional layer is passed to the next point-by-point convolutional layer through the deep convolutional inversion layer for the inversion operation, and it is passed to the next convolutional block along with the feature information of the original feature map through the deep convolutional fusion layer for the feature summation, and then to the next point-by-point convolutional layer, and so on. To prevent gradient dispersion, an inverted residual structure is added to the convolution block. The whole network is optimized by cross entropy loss function until convergence. Finally, the global average pooling layer and point-by-point convolutional layer are used to achieve the number of classification requirements, and the input samples are classified by Softmax classifier.

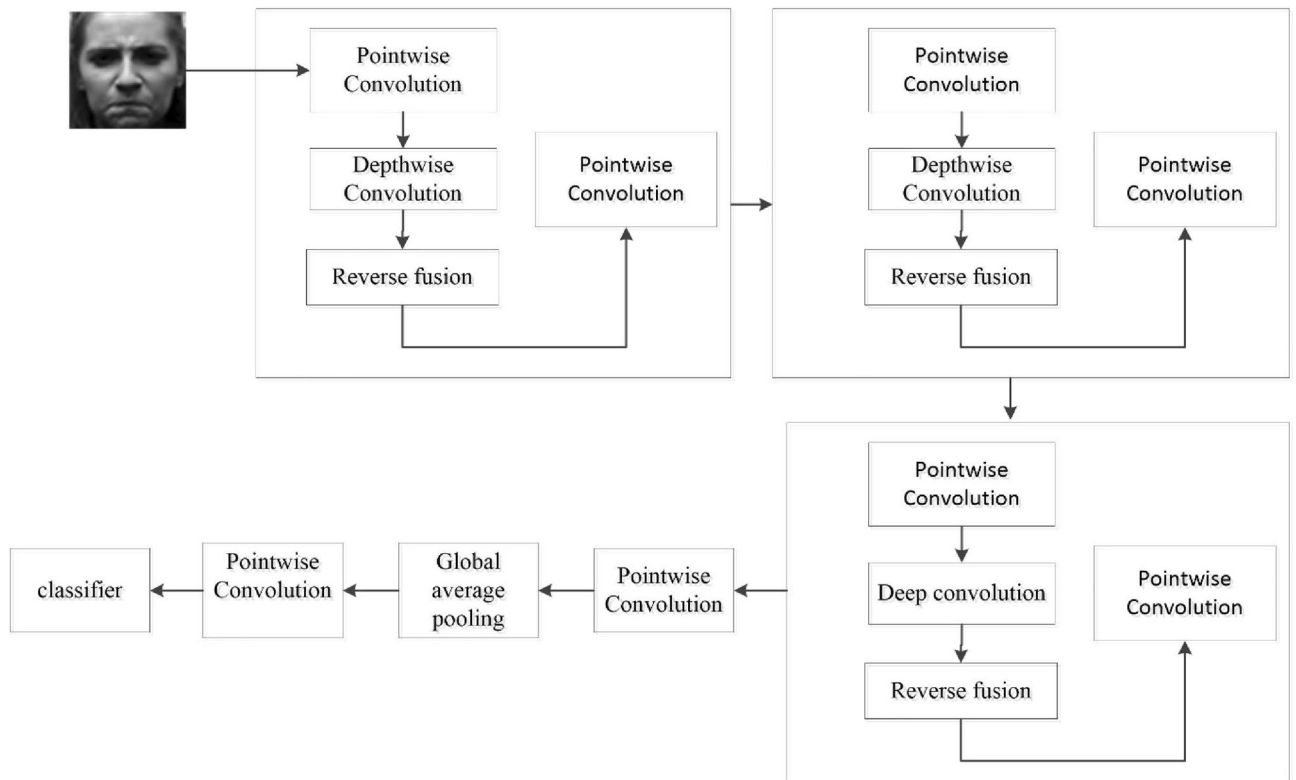


Figure 3. Schematic diagram of reverse fusion operation.

Attention mechanism

The attentional mechanism is a signal processing mechanism discovered by some scientists in the 1990s while studying human vision. Practitioners in the field of artificial intelligence have introduced this mechanism into some models with success. Nowadays, the attention mechanism has become one of the most widely used “components” in the field of deep learning, especially in the field of natural language processing. The models or structures such as BERT, GPT, Transformer, etc., which have received a lot of exposure in the past two years, all employ attention mechanisms.

Attention mechanisms can be categorized into attention based on spatial relationships and attention based on channel data. The channel-based attention network is to consider the detection target as a single individual, put the addressing target of attention on the addressing of some features of a single individual, and synthesize the feature information of different locations of the individual to give reasonable prediction results. The SE (Squeeze-and-Excitation Networks) attention mechanism is one of the channel attention mechanisms. Essentially, SENET is doing attention on the channel dimension, and this attention mechanism allows the model to pay more attention to the most informative channel features, while suppressing those unimportant or ineffective channel features to train the model for better results. Moreover, the network structure of SENET is relatively simple and can be easily integrated into various neural network structures to improve the performance of the network, and its structure is shown in Fig. 4.

SENET contains a compression operation (Squeeze), an excitation operation (Excitation) and a weight calculation (Scale). For a given input image, the initial extraction of features is accomplished by a convolutional neural network. And then SENET is used for feature weighting calculation, and the calculation process is described as follows, in Eq. 6, F_{sq} denotes global average pooling, while U_c denotes the c -th two-dimensional matrix in U and the subscript c denotes channel. $U_c(i, j)$ represents the value in the c -th two-dimensional matrix in U , where $1 \leq i \leq H, 1 \leq j \leq w$.

H and W denote the height and width of the feature map, and U_c denotes the number of channels of the feature map. δ denotes the relu activation function, σ denotes the sigmoid function, W_1 and W_2 denote the two fully-connected operations, A_c and the weight values of the input features obtained through the SE module denote the outputs obtained through the attention mechanism, U'_c which are used as inputs for the subsequent computations.

$$Z_c = F_{sq}(U_c) = \frac{1}{H \times w} \sum_{i=1}^H \sum_{j=1}^w U_c(i, j) \quad (6)$$

$$A_c = F_{ex}(Z_c, W) = \sigma(W_2 \delta(W_1 Z_c)) \quad (7)$$

$$U'_c = A_c \cdot U_c \quad (8)$$

Activation function

The Relu6 activation function is used in mobilenetv2 network, the relu6 activation function adds a nonlinear constraint to the positive semiaxis of Relu will be greater than six inputs all the output to six to prevent the gradient explosion and the calculation is simple, its expression is as follows:

$$f(x) = \min(\max(0, x), 6) \quad (9)$$

However, relu6 still has the shortcoming that the negative gradient information disappears, and the mean and variance are not 0. In order to improve this problem, the SELU activation function is introduced in this paper.

SELU allows the input to become a stationary distribution after a certain number of layers, and the positive semiaxis of SELU is greater than 1, which allows it to increase when the variance is too small, but at the same time prevents the gradient from vanishing. So, the activation function then has an immobility point, and the output of each layer when the network is deeper has a mean of 0 and a variance of 1. β and α are two hyperparameters, and studies have shown that the value of β is usually taken to be 1.05, and the value of α is usually taken to be 1.67. Equation 10 is the SELU function expression. Figure 5 shows the relu6 and SELU function image and its derivative image, where Fig. 5a is a schematic diagram of the SELU function and relu6 function image, when the input is less than 0, the relu6 function output is all 0, at this time, the activation function loses its role, and the

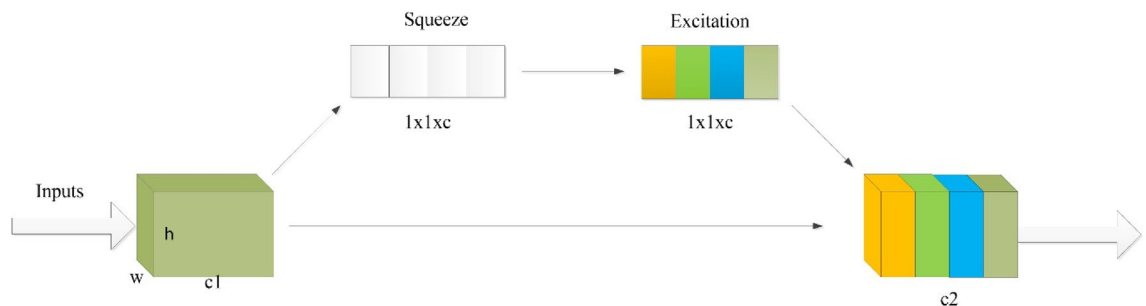


Figure 4. Schematic diagram of the attention mechanism.

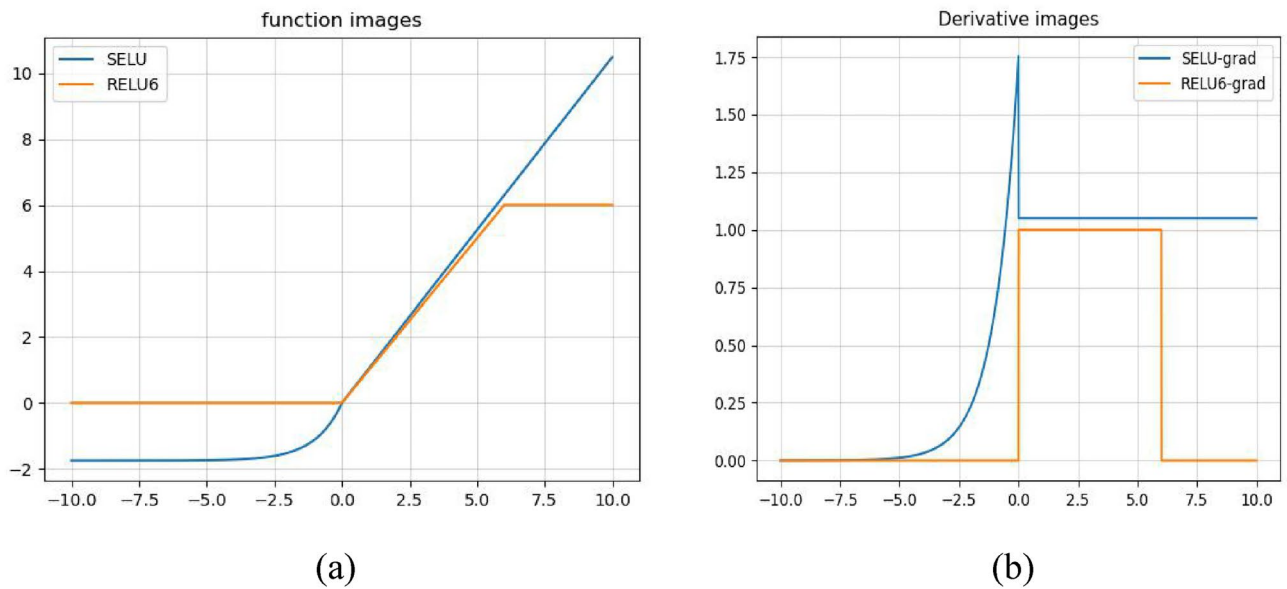


Figure 5. relu6 with SELU function and its derivative image.

SELU function is still able to maintain the negative information, and Fig. 5b is a derivative image of the SELU function and the relu6 function, and when backpropagation is performed When the value is 0, if the value is 0, relu6 gradient information is all 0, the gradient is not being updated, SELU function can still backpropagate and continue to learn the features.

$$f(x) = \beta \times \begin{cases} \alpha(e^x - 1)x < 0 \\ xx \geq 0 \end{cases} \quad (10)$$

Improved mobilenetv2 network

The core idea of MobilenetV2 network is to adopt inverted residual structure and linear bottleneck on the basis of MobilenetV1. In order to avoid direct deletion of negative information by the relu6 activation function and to enhance the ability of information extraction, the reverse fusion mechanism and the attention mechanism are integrated into MobilenetV2 network. At the same time, the relu6 activation function is modified to SELU activation function. Figure 6 shows the schematic diagram of the improved inverted residual structure. Table 1 shows the network structure of MobilenetV2.

The input features are first upsampled by a 1*1 convolution, then by a 3*3 convolution, and finally downsampled by a 1*1 convolution layer. Stride denotes the step size, which is used to down sample the feature layer size reduction. Down sampling is only used in the first inverted residual structure of each bottleneck layer (no residual join is used at this point). The inverted residual structure does not use residual connections when stride=2 or when the input and output channels are not the same. Table 1 shows the network structure table of mobilenetv2, where t denotes the expansion factor and n denotes the number of repetitions.

An image of size $3 \times 48 \times 48$ is input into the network model. The input image first passes through a convolutional layer composed of 3×3 convolutional kernels, then through the main framework of the network stacked by the inverted residual module, and then through the global average pooling layer Global Average Pooling(GAP). Finally, a fully connected layer is used to predict the final score of each expression, and the expression with the highest score is the predicted expression. The feature maps are normalized by the Batch Normalization layer after the convolution operation in the convolution layer. Because reducing the width factor can significantly reduce the number of network model parameters and the loss of network accuracy can be controlled. In this paper, we make $\alpha=0.75$ and reduce the highest output dimension of the network to 320, and improve the network accuracy by other improvement measures.

Experimental method and results analysis

Experimental environment

The experiments in this paper on PC are based on pytorch1.8.0 deep learning framework and the programming language is python 3.9 on WINDOWS10 64-bit operating system. The hardware platform is Intel(R) Core(TM) i7-8750H CPU @ 2.20 GHz 2.21 GHz. GPU: NVIDIA GeForce GTX 1050 Ti, CUDA 12.1.

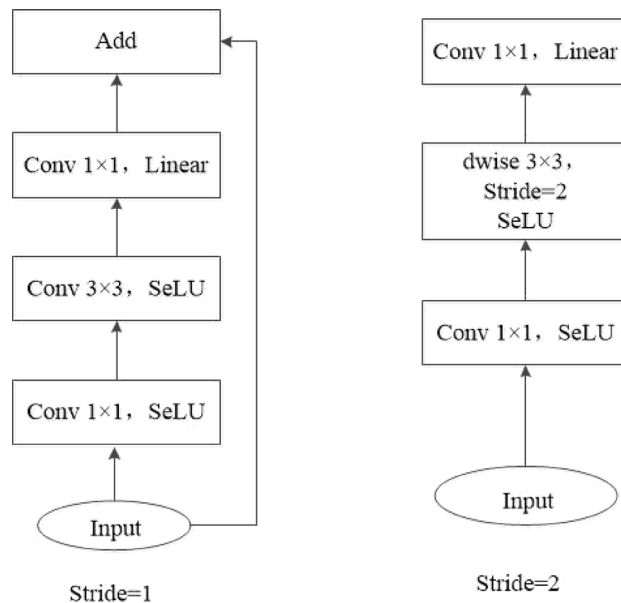


Figure 6. Schematic diagram of inverted residual structure.

Number	t	Inputs	Stride	n	Outputs	Framework	Reverse fusion	SNet
1	-	$3 \times 48 \times 48$	2	6	$16 \times 24 \times 24$	Conv2d 3×3		
2	2	$16 \times 24 \times 24$	2	2	$24 \times 12 \times 12$	block	√	√
3	6	$24 \times 12 \times 12$	1	3	$32 \times 6 \times 6$	block	√	
4	6	$32 \times 6 \times 6$	1	3	$64 \times 3 \times 3$	block	√	
5	6	$64 \times 3 \times 3$	1	4	$96 \times 3 \times 3$	block	√	
6	6	$96 \times 3 \times 3$	1	2	$160 \times 2 \times 2$	block	√	√
7	6	$160 \times 2 \times 2$	1	3	$320 \times 2 \times 2$	block	√	
8	-	$320 \times 2 \times 2$	1	1	$320 \times 1 \times 1$	Avgpool2d		
9	-	$320 \times 1 \times 1$	1	1	k	Conv2d 1×1		

Table 1. MobilenetV2 network structure table.

Results and discussion

Analysis of experimental results for the FER2013 dataset

The model is tested on FER2013 dataset, in the experiment the SGD optimizer is used to optimize the loss, the momentum (Momentum) is set to 0.9, the initial learning rate is 0.01, the learning rate weighting factor is $5e-4$, epoch is 350, and the batch size is 32. The training set contains a total of 28,709 images, and the public test set (3589 images) is used to adjust the weighting parameters and finally tested on the private test set (3589 images in total). The following Fig. 7 shows the confusion matrix of the network model on the FER2013 dataset.

As can be seen in Fig. 7, the accuracy of the proposed model in this paper is 68.626% on FER dataset. The improvement compared to mobilenetv2 indicates that the feature extraction ability of the model is enhanced by reverse fusion, introduction of the attention mechanism and adoption of the SELU activation function for more accurate recognition. Among them, the recognition rate is higher on happy, surprise, disgust and neutral reaching 89%, 81%, 71%, 70% respectively. Anger, fear, and sadness are prone to misjudgment due to more similarities in their expressions, resulting in lower recognition rates.

Analysis of experimental results for the CK + dataset

The network model in this paper was trained and tested on the CK + dataset, and the weights and biases were randomly initialized during the training process, with a batch size of 20, a number of iterations of 150, an initial learning rate of 0.01, and a stochastic gradient descent optimizer with a momentum factor of 0.9. The model training process was performed using the training set (882 frames) of CK + and tested on the test set (99 frames). The network model is pre-processed by randomly flipping the samples before parameter training on the training dataset. The confusion matrix of the network model on the CK + dataset is shown in Fig. 8.

The model proposed in this paper has an accuracy of 95.960% on the CK + dataset and achieves 100% accuracy on disgust, fear, happy, sad, and surprise, 83% recognition accuracy contempt, and 75% recognition on angry.

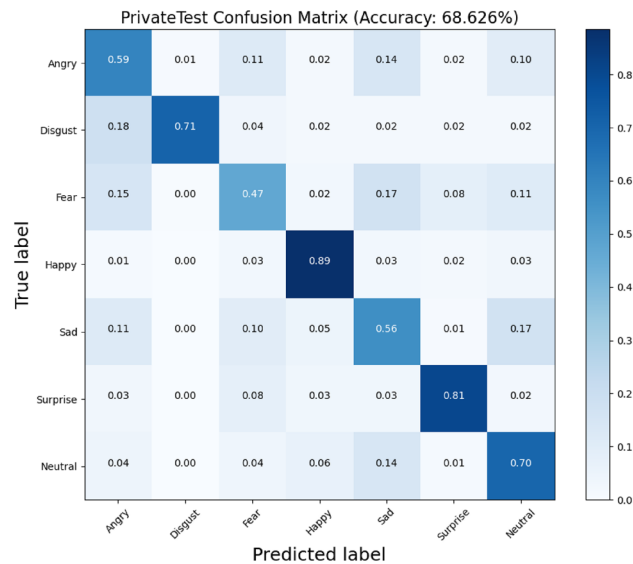


Figure 7. Confusion matrix plot on the FER2013 dataset.

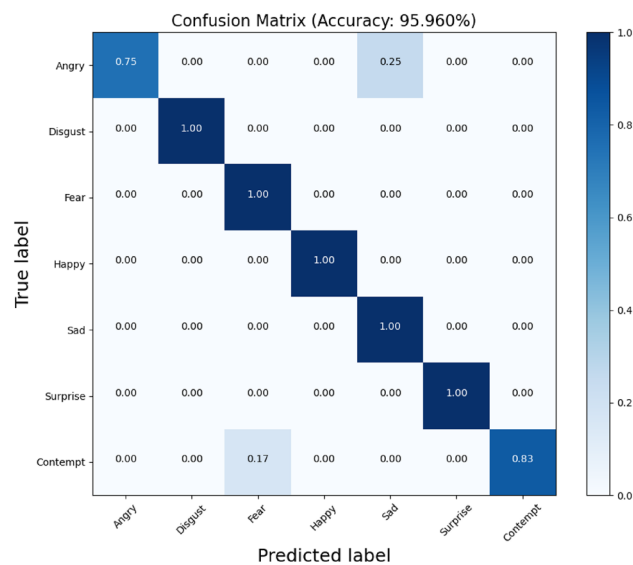


Figure 8. Confusion matrix on the CK+ dataset.

All have achieved better results. There is a significant improvement relative to mobilenetv2, indicating that the feature extraction ability of the model is enhanced by reverse fusion, introduction of the attention mechanism and adoption of the SELU activation function, and the recognition is more accurate.

Ablation study

In this paper, methods such as adding attention mechanism, reverse fusion mechanism and SELU activation function are used to strengthen the feature extraction ability of the network, and at the same time, the compression of the width factor is used to reduce the number of model parameters. This paper conducts comparative ablation study on two datasets, FER2013 and CK+ to analyze the impact of different improvements in on the experimental results. The number of model parameters in this paper is only 3.26 M, which is only 16.2% of the base model. However, the performance is improved by 0.72% and 5.10% on the two datasets, especially on the CK+ dataset. Both the reverse fusion mechanism and the attention module can improve the model performance, the SELU activation function has no effect on the number of model parameters, and the three approaches achieve the effect of reducing the number of parameters and improving the model accuracy when acting on the mobilenetv2 base model. Table 2 shows the comparison of the accuracy of different models using different parts on the dataset.

NET	Params Size/MB	Accuracy rate /%	
		FER2013	CK +
Mobilenetv2	20.10	67.90	89.82
Mobilenetv2(0.75)	3.33	66.99	86.82
Mobilenetv2(0.75) + SE	3.45	67.48	90.91
Mobilenetv2(0.75) + RFF	3.33	68.22	91.92
Ours	3.26	68.62	95.96

Table 2. Results of ablation experiments comparing FER2013 with CK + dataset.

Comparison experiment results and analysis

To compare the advancedness of the proposed network model, this section will compare it with the current mainstream network model. As shown in Table 3, there is a slight difference in accuracy between the network model in this paper and the traditional network model, but there is a significant decrease in the number of parameters, which is only 3.6% and 4.5% of that of VGG19 and Res-Net50. On the FER2013 dataset, the accuracy of the base model used in this paper decreases compared with the traditional deep neural network. On the CK + dataset, the accuracy is equal to that of ResNet50 and 1.32% higher than that of VGG19. Among the lightweight network models, the model in this paper has different degrees of improvement in accuracy and number of parameters, and the improvement is more obvious in CK +. On the FER2013 dataset, the accuracy is improved by 1.06% and 1.46%, and the number of parameters is reduced by 16.84 MB and 18.42 MB compared with the same series of MobileNetV2 and MobileNetV3, respectively. Compared with Xception, a lightweight network, the accuracy increased by 1.7% and the number of parameters decreased by 0.79 MB. On the CK + dataset, the accuracy is improved by 6.14% and 7.09% compared with the same series of Mobile-NetV2 and MobileNetV3. Compared with other lightweight networks, Xception, the accuracy is improved by 3%.

By comparing with the current mainstream deep convolutional neural networks and lightweight networks on the two datasets, it can be seen that the model in this paper greatly reduces the number of parameters with a small amount of accuracy loss, and the comprehensive performance is higher than that of the current mainstream lightweight networks. The main reasons are.

- (1) The adjustment of the depthwise separable convolution and width factor makes the number of parameters of the model decrease significantly.
- (2) The reverse fusion mechanism is introduced to reduce the loss of feature information in the convolutional layer.
- (3) The SE mechanism is introduced, and the attention mechanism is integrated in the dimension of the channel, which significantly enhances the ability of extracting effective feature information and weakens the extraction of invalid information.

Real-time testing on mobile

In order to verify the real-time performance of this paper's model on the mobile side, this paper also compares the real-time performance of I-MobileNetV2 with MobileNetV1 and MobileNetV2 models on Huawei p60 cell phone on the mobile side, and selects an image of a typical expression of each of the seven expressions on the CK + dataset to make predictions, and the results of the predictions made by the various models after 1,000 times are shown in Table 4.

In Table 4, the real-time test of ours network is much better than that of MobileNetV1 and MobileNetV2, indicating that the model in this paper has better real-time performance compared to MobileNetV1 and MobileNetV2 models, and the model in this paper reduces the network parameters by adjusting the width factor, and also improves the real-time performance.

NET	Params Size/MB	Accuracy Rate/%	
		FER2013	CK +
Vgg	89.59	70.30	94.64
Resnet50	72.31	72.86	95.96
Mobilenetv2	20.10	67.90	90.90
Mobilenetv3 ²⁹	21.50	67.50	88.87
Xception ³⁰	4.35	67.26	92.96
Ours	3.26	68.62	95.96

Table 3. Accuracy of different models for expression recognition on FER2013 and CK + dataset.

model	Average prediction time/ms
MobileNetV1	221
MobileNetV2	300
ours	78

Table 4. Performance comparison of different models on mobile devices.

Conclusion

In this paper, a lightweight facial expression recognition network model with feature fusion channel attention mechanism is proposed to solve the problems of complex structure of traditional convolutional neural network, which leads to excessive parameterization, and insufficient extraction of feature information of face expression. The network model is based on the MobileNetV2 network, the width factor and the overall network dimension are compressed, the feature information is fused, and the efficient channel attention module is embedded to recognize facial expressions. The accuracy of the proposed network model on FER2013 and CK+ datasets reaches 68.62% and 95.96% respectively, which is 0.72% and 6.14% higher than that of MobileNetV2, and the number of parameters decreases by 83.79%. The experimental results show that the network model in this paper shows better performance in terms of accuracy and number of parameters compared with many other convolutional neural network models, which verifies the effectiveness of the network model improvement measures. In the subsequent more in-depth research, the lightweight multi-scale convolution can be used for feature extraction and feature fusion to solve the problems of insufficient ability of single-scale convolutional kernel to extract feature information and limited sensory field.

Data availability

FER2013 and the “CK+ database” are publicly available and can be downloaded at <https://www.kaggle.com/datasets/msambare/fer2013>. <https://www.kaggle.com/datasets/shawon10/ckplus>.

Received: 20 January 2024; Accepted: 2 April 2024

Published online: 07 April 2024

References

1. Adyapady, R. R. & Annappa, B. A comprehensive review of facial expression recognition techniques. *Multimed. Syst.* **29**(1), 73–103 (2023).
2. Ge, H. *et al.* Facial expression recognition based on deep learning. *Comput. Methods Progr. Biomed.* **215**, 106621 (2022).
3. Savchenko A V. Personalized frame-level facial expression recognition in video. International Conference on Pattern Recognition and Artificial Intelligence. *Cham: Springer International Publishing*. 447–458(2022).
4. Feighelstein, M. *et al.* Automated recognition of pain in cats. *Sci. Rep.* **12**(1), 9575 (2022).
5. Jia, H., Xiao, Z. & Ji, P. Real-time fatigue driving detection system based on multi-module fusion. *Comput. Graph.* **108**, 22–33 (2022).
6. Sivaiah, B. *et al.* FL-CapsNet: facial localization augmented capsule network for human emotion recognition. *SIViP* **17**(4), 1705–1713 (2023).
7. Pan, B. *et al.* A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing* **561**, 126866 (2023).
8. Nan, F. *et al.* Feature super-resolution based facial expression recognition for multi-scale low-resolution images. *Knowl.-Based Syst.* **236**, 107678 (2022).
9. Canal, F. Z. *et al.* A survey on facial emotion recognition techniques: a state-of-the-art literature review. *Inf. Sci.* **582**, 593–617 (2022).
10. Vu, H. N., Nguyen, M. H. & Pham, C. Masked face recognition with convolutional neural networks and local binary patterns. *Appl. Intell.* **52**(5), 5497–5512 (2022).
11. Arun, M. & Arivazhagan, S. A unified feature descriptor for generic character recognition based on zoning and histogram of gradients. *Neural Comput. Appl.* **34**(14), 12223–12234 (2022).
12. Burger W, Burge M J. Scale-invariant feature transform (SIFT). *Digital Image Processing: An Algorithmic Introduction*. *Cham: Springer International Publishing*. 709–763(2022).
13. Chen, X. *et al.* Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **79**, 102444 (2022).
14. Sharifani, K. *et al.* Operating machine learning across natural language processing techniques for improvement of fabricated news model. *Int. J. Sci. Inform. Syst. Res.* **12**(9), 20–44 (2022).
15. Bhardwaj, V. *et al.* Automatic speech recognition (asr) systems for children: a systematic literature review. *Appl. Sci.* **12**(9), 4419 (2022).
16. Tadiparthi, P. K., Bugatha, S. & Bheemavarapu, P. K. A review of foreground segmentation based on convolutional neural networks. *Int. J. Adv. Comput. Sci. Appl.* <https://doi.org/10.14569/IJACSA.2022.0130951> (2022).
17. Helaly, R. *et al.* DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. *Signal, Image Video Process.* **17**, 1–14 (2023).
18. Kayadibi, I. *et al.* An eye state recognition system using transfer learning: alexnet-based deep convolutional neural network. *Int. J. Comput. Intell. Syst.* **15**(1), 49 (2022).
19. Yen, C. T. & Li, K. H. Discussions of different deep transfer learning models for emotion recognitions. *IEEE Access.* **10**, 102860–102875 (2022).
20. Ghosh T, Banna M H A, Nahian M J A, *et al.* A privacy-preserving federated-mobilenet for facial expression detection from images. In: *International Conference on Applied Intelligence and Informatics*. *Cham: Springer Nature Switzerland*. 277–292(2022).
21. Banerjee, A. *et al.* Training and profiling a pediatric facial expression classifier for children on mobile devices: machine learning study. *JMIR Form. Res.* **7**, e39917 (2022).

22. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520(2018).
23. Gulzar, Y. Fruit image classification model based on MobileNetV2 with deep transfer learning technique. *Sustainability* **15**(3), 1906 (2023).
24. Pang, Y., Yu, W., Zhang, Y., Xuan, C. & Wu, P. Sheep face recognition and classification based on an improved MobilenetV2 neural network. *Int. J. Adv. Robot. Syst.* <https://doi.org/10.1177/17298806231152969> (2023).
25. Kumar, B. A. & Bansal, M. Face mask detection on photo and real-time video images using caffe-mobilenetV2 transfer learning. *Appl. Sci.* **13**, 935. <https://doi.org/10.3390/app13020935> (2023).
26. Jiang, M. & Yin, S. Facial expression recognition based on convolutional block attention module and multi-feature fusion. *Int. J. Comput. Vision Robot.* **13**(1), 21–37 (2023).
27. Khaireddin Y, Chen Z. Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv.2105.03588* (2021).
28. Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. 2010: 94–101(2010).
29. Howard, A. et al. Searching for mobilenetv3. *Proc. IEEE/CVF Int. Conf. Comput. Vision.* **2019**, 1314–1324 (2019).
30. Fran C. Deep learning with depth wise separable convolutions. In: *IEEE conference on computer vision and pattern recognition (CVPR)*. (2017).

Author contributions

HZ contributed to the conception of the study; QZ performed the experiment;MZ contributed significantly to analysis and manuscript preparation;SX performed the data analyses and wrote the manuscript;RM helped perform the analysis with constructive discussions.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024