Journal of Advanced Computing and Intelligent Systems (JACIS)
https://journal.chainknowpub.com/index.php/JACIS/index
ISSN Online: –
ISSN Print: –

An Improved Xception Model with Multi–scale Feature Fusion and Convolutional Block Attention Module (CBAM) for AI–Driven Facial Expression Recognition

Shuzlina Abdul–Rahman

Universiti Teknologi MARA Cawangan Terengganu, Kuala Terengganu, Malaysia

Abstract: The continuous development of deep learning (DL) has significantly impacted computer vision tasks, particularly with the convolutional neural network (CNN) Xception. In facial expression recognition, the implementation of the deep separable CNN Xception has greatly improved the neural network㊄s representation ability, thereby improving the facial recognition rates. However, there are two critical issues in facial feature extraction. The first issue is that the single–scale expression features fail to capture the rich facial expression information adequately while the second issue is that expression features are not evenly distributed across the entire face image. The first issue is addressed by designing a multi–scale and multi–channel recognition method to extract these diverse features. Meanwhile, the Convolutional Block Attention Module (CBAM) is incorporated into the multi–scale and multi–channel convolutional neural network. This article chooses Xception as the basic network, uses multi–scale depth separable modules with convolution kernels of 3 ˙ 3, 5 ˙ 5 and 7 ˙ 7 to extract richer expression features, and performs multi–channel convolution feature fusion to add attention mechanism module. The proposed model achieved a recognition accuracy of 94.35% on the dataset FER2013, which is 1.6% higher than unimproved Xcepiton network， and the results were more obvious compared to the other five network models, verifying the effectiveness of the improvement measures.The proposed method demonstrates significant engineering application value and can be widely utilized in areas such as telemedicine, smart education, and autonomous driving.

Keywords: Xception; multi–scale depth separable module; multi–channel; feature fusion; CBAM

1.Introduction

Over the past 30 years, facial expression recognition(FER) has attracted great attention from many researchers and scientists, especially the problem of emotion classification. Based on cross–cultural research, Voleti, S., et al. [1]proposed to classify human expressions into six basic facial expressions (anger, disgust, fear, happy, sad and surprise), and proposed that basic human facial expressions will not be different due to cultural differences. How to distinguish each facial expression, extract its unique features, and classify them is a problem that researchers consider.

Porusniuc, G.C., et al.[2] shows the traditional FER process mainly includes steps such as face detection, key point location, facial feature extraction, expression feature optimization, feature classification, and expression recognition. The FER process based on deep learning mainly includes steps such as face detection, data preprocessing, feature extraction, and expression recognition. Zhang, Y., et al.[3] defined convolutional neural networks (CNN) is a type of feedforward neural network(FNN) that includes convolution calculations and deep structure. It is one of the representative algorithms for deep learning(DP), it improved recognition accuracy than traditional methods. Compared with traditional CNN methods, deep CNNs[4] show higher robustness and effectiveness, with powerful capabilities for extracting, learning, and classifying features, and are very effective in identifying and discerning subtle changes in facial expressions. With the real-world needs of computer vision and the high computational requirements of CNN networks, almost all deep CNNs for facial expression recognition face two problems.

One problem[5] is that the network depth of CNN is getting deeper and deeper, and the number of model parameters is getting larger and larger. How to reduce the number of trainable parameters, how to improve the structure of CNNs, and how to adjust the network depth and width are becoming increasingly difficult. Francois Chollet [6]proposed the introduction of depthwise separable convolution based on Inception v3, which improved the recognition effect of the network model without increasing the complexity of the network. It is considered a milestone in the lightweight CNN.

Another problem is the difficulty in extracting and classifying facial expression features. On one side, single-scale features[7] cannot describe rich facial expression information, on other hand, expression features are not evenly distributed on facial images. The usual approach to this problem is to introduce attention mechanisms and multi-scale feature fusion, thereby truly improving the recognition rate of facial expression recognition. Xusong Luo[8] designed a multi-scale local feature fusion network to improve the performance of FER in actual application scenarios. Wanzhao Li[9] proposed a framework combines global features with several different local key features to consider the multiple labels of expressions embodied in many facial action units. Hao Lin[10] proposed an improved Xception with dual attention mechanism and feature fusion for face forgery detection, He et al.[11] introduced the squeeze and excitation (SE) module to enhance the channel relationship by recalibrating the importance of each channel.

In addition, Zhao et al.[12] obtained robust and diverse information by utilizing multi-scale features, and proposed a lightweight fully convolutional attention network LANMSFF, which uses two novel modules of large-scale attention and point-by-point feature selection, so that the extracted features can provide robust and diverse representation information in multi-view situations.

However, these methods simply fuse features from different scales and assign the same weight to these features. However, utilizing all features from different angles without considering their importance may have a negative impact on recognition accuracy..

Our contributions can be summarized as follows: We designed an improved Xception model with multi-scale feature fusion and CBAM for facial expression recognition.

(1) In the entry flow, use convolution kernels of 3 ˙ 3, 5 ˙ 5, and 7 ˙ 7 depth separable convolution modules to extract features from facial expression images, and perform feature splicing and fusion to expand the receptive field and obtain richer facial feature information.

(2) In the middle flow, a convolutional attention module embedding channel attention and spatial attention in series is proposed, so that the network model can focus on extracting key feature information in channels and spaces, effectively enhancing the representation of the network model ability.

(3) The proposed model achieved a recognition accuracy of 96.45% on the main dataset FER2013. Can be applied to facial expression recognition on most devices.

2.Related work

In this section, we give a brief literature review of this paper, including prior works on lightweight network architecture and attention mechanism.

2.1 Lightweight network

Xception network is another improvement to Inception V3 proposed by Google Company after Inception V3. It mainly uses depth-separable convolution to replace the classical convolution operation in Inception V3, and also refers to the residual separable convolution of ResNet. And the structural unit of the network can increase the number of layers of the model while reducing the amount of parameters, which not only reduces the storage space, but also enhances the expressive ability of the model[13].

The architecture diagram of Xception is shown in Figure 1, which mainly includes 36 layers of convolution. it can be divided into 3 flows, namely Entry flow, Middle flow, and Exit flow. it is divided into 14 blocks, including 4 blocks in Entry flow, 8 blocks in Middle flow, and 2 blocks in Exit flow.

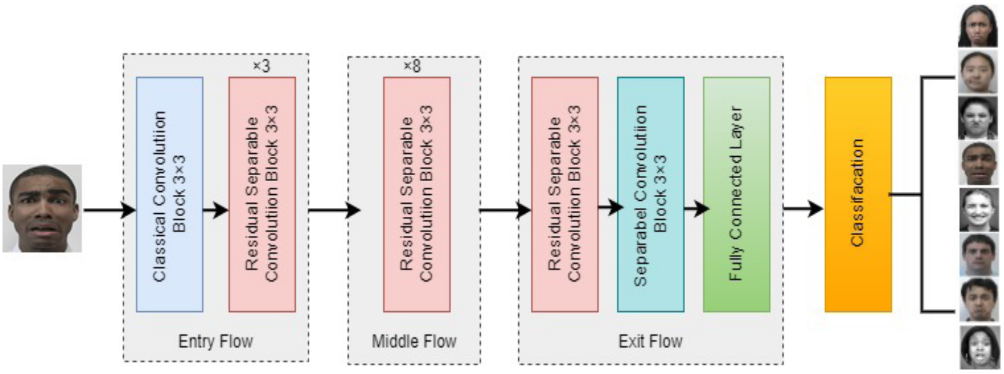
×3
Classical Convolutiion Block 3×3
Residual Separable Convolution Block 3×3
Entry Flow
×8
Residual Separable Convolution Block 3×3
Middle Flow
Residual Separable Convolution Block 3×3
Separabel Convolution Block 3×3
Fully Connected Layer
Exit Flow
Classifacation

Figure 1 The architecture diagram of Xception

The Xception network simplifies the network in Inception structure, retains all convolution branches and merges and then increases the number of convolution branches of the output channel of the convolution. In fact, each convolution acts as the feature map containing one channel, this is the depth-separable convolution module. The Xception model is composed of some linear stacks of depth-separable convolution block and ResNet-like residual separable convolution blocks [14].The residual separable convolution block is shown in Figure 2.

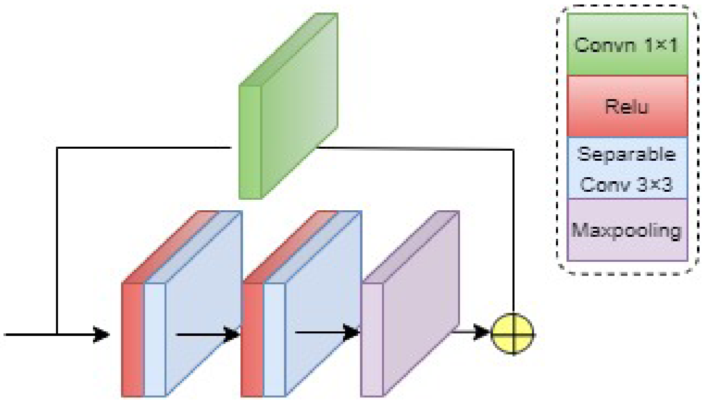

Figure 2 residual separable convolution block

2.2 Multi-scale feature fusion

The receptive field of the convolutional neural network determines the scale of the extracted image features and the area that each pixel of the feature map can perceive. The size of the convolution kernel will directly affect the size of the receptive field[15]. When the size of the convolution kernel in the convolution layer is set too small, the convolutional neural network will capture too many local features and ignore the global features; when the convolution kernel is set too large, it will not be able to capture detailed information, and it will also cause the problem of feature information redundancy.

Multi-scale feature fusion[8] is to extract multi-scale features and fuse them by inputting an image using convolution kernels of different sizes. There are usually two types of feature fusion operations: feature map addition and feature map concatenation (Concat). The Add operation adds all elements of the two feature maps one by one, so it is necessary to ensure that the w, h and c of the two feature maps are the same, as shown in Figure 3.

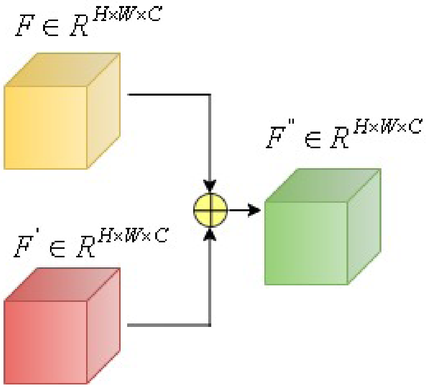
$F \in R^{H\times W\times C}$
$F'' \in R^{H\times W\times C}$
$F' \in R^{H\times W\times C}$

Figure 3 Feature map addition operation

The concatenation operation concatenates two feature maps by channel, so it is necessary to ensure that w and h of the two feature maps are the same. The number of channels of the concatenated feature map increases, but w and h do not change, as shown in Figure 4[16].

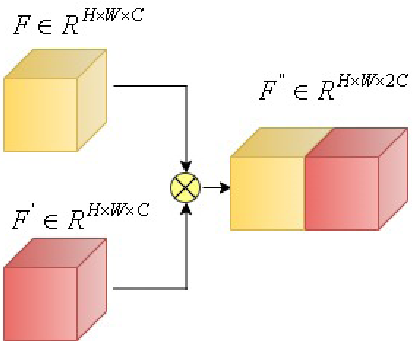
$F \in R^{H\times W\times C}$
$F'' \in R^{H\times W\times 2C}$
$F' \in R^{H\times W\times C}$

Figure 4 Feature map concatenation operation

2.3 Multi-channel convolution

Multi-channel convolution[17] is a convolution operation commonly used in convolutional neural networks, which can effectively extract multiple features of input data. Multi-channel convolution can process multi-channel data such as color images well, extract features of different channels and combine them effectively. Multi-channel convolution uses multiple convolution kernels to perform convolution operations on each channel of the input data, and then adds the convolution results of each channel to obtain the final output[18].If input the image $\boldsymbol{F} \in \boldsymbol{R}^{8\times8\times3}$,has 4 filters, after the multi-channel convolution operation, the output image becomes $\boldsymbol{F}' \in \boldsymbol{R}^{6\times6\times4}$.

2.4 Attention Mechanism

Attention Mechanism is an important technology in the field of deep learning, which allows the model to focus on important parts of the input data, thereby improving overall performance and efficiency. The attention mechanism mimics the human attention process, allowing the model to focus on key information in the input data while ignoring less relevant parts. In traditional convolutional neural networks, the output of each neuron depends on the output of all neurons in the previous layer. After the attention mechanism is introduced, the output of each neuron not only depends on the output of all neurons in the previous layer. Different weights can also be assigned to different neurons. By assigning different weights to different parts of the input data, the model is able to identify the most important information.

Attention mechanisms have proven to be very effective in image classification and image segmentation. For example, SENet is a very good attention mechanism network that simply compresses each 2D feature map to effectively build the interdependencies between channels. CBAM takes this idea further by introducing spatial information encoding using convolutions with large-sized kernels. Later, more and more researchers designed many attention networks by using different spatial attention mechanisms, such as GENet, GALA, AA and TA, etc.

Recently, self-attention networks are very popular and have very powerful ability to construct spatial or channel attention. Typical examples include NLNet, GCNet, A2Net, SCNet, GSoP-Net or CCNet, all of which exploit non-local mechanisms to capture different types of spatial information. However, networks incorporating self-attention mechanisms require high-configuration computer hardware resources and are usually suitable for large models but not suitable for lightweight convolutional neural networks.

To improve lightweight neural networks, we should consider a more effective method to capture position information and channel relationships to enhance the feature representation of the network. We should focus on attention methods with lightweight properties, such as SENet, CBAM, and TA.

The Squeeze-and-Excitation(SE) module is an attention mechanism proposed by Hu et al..It aims to improve the performance of convolutional neural networks by recalibrating the channel weights of feature maps. The SE module is composed of two parts, Squeeze and Excitation, connected in series.Assuming that the input feature is $\boldsymbol{M}^{\mathrm{H}\times\mathrm{W}\times\boldsymbol{C}}$, in Squeeze, first perform global average pooling on the input feature map according to different channels, as shown in Formula 1:

$$\boldsymbol{M_2}^{1\times1\times\boldsymbol{C}}=\boldsymbol{GlobalAvgPool(M}^{\mathrm{H}\times\mathrm{W}\times\boldsymbol{C}}\boldsymbol{)} \qquad (1)$$

Compress the input feature $\mathrm{H}\times\mathrm{W}\times\boldsymbol{C}$ to $1\times1\times\boldsymbol{C}$, then enter Excitation and connect the compressed features through two fully connected layers to obtain the weights on each channel. Finally, multiply the output features and weights of the convolution module by channel to obtain the output features that incorporate channel attention.

Convolutional Block Attention Module(CBAM)[19] is an attention mechanism module for convolutional neural networks, which aims to enhance feature representation capabilities by paying attention to the channel dimension and spatial dimension of the feature map respectively.

3 Proposed model

The proposed method also includes three flows including entry flow,middle flow and exit flow like Xception network. The Architecture of The proposed model is exactly shown in Figure 5.

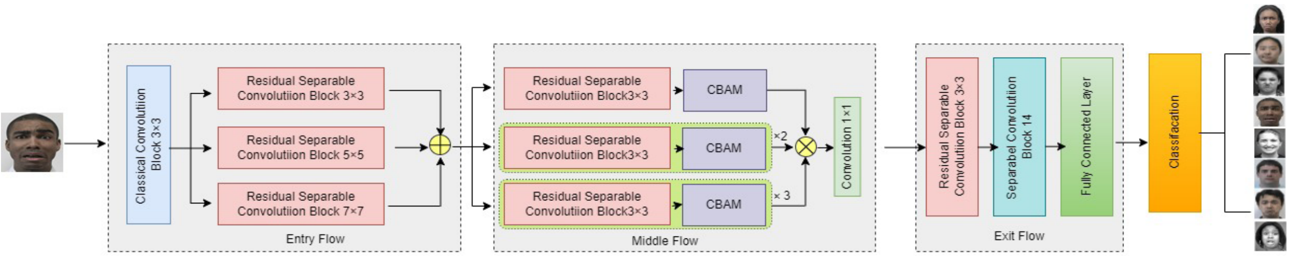
Classical Convolutiion Block 3×3
Residual Separable Convolutiion Block 3×3
Residual Separable Convolutiion Block 5×5
Residual Separable Convolutiion Block 7×7
Entry Flow
Residual Separable Convolutiion Block3×3
CBAM
Residual Separable Convolutiion Block3×3
CBAM
×2
Residual Separable Convolutiion Block3×3
CBAM
× 3
Convolution 1×1
Middle Flow
Residual Separable Convolutiion Block 3×3
Separabel Convolutiion Block 14
Fully Connected Layer
Exit Flow
Classifacation

Figure 5 The Architecture of The proposed model

A.Entry flow

The number of blocks in the proposed entry flow is exactly the same as that in original Xception model (refer to Figure 1), which consists of a classical convolution block and 3 residual separable convolution blocks. Firstly, two 3 ˊ 3 convolution blocks are used to extract shallow features, and then three convolution blocks of different scales, 3 ˊ 3, 5 ˊ 5, and 7 ˊ 7, are used to extract facial expression features of different sensory fields.

B.Middle flow

The middle flow in the original Xception serializes 8 residual separable convolution blocks on the same branch. The proposed method reduces 8 residual separable convolution blocks to 6 residual separable convolution blocks, and distributes them to three parallel branches according to 1, 2, and 3 blocks. In this way, different high-dimensional semantic feature images can be extracted using different levels of convolution. At the same time, CBAM is introduced after each separabelConv block to enhance the representation ability of features and refine intermediate features. Finally, 1x1 convolution is used to merge the reorganized multi-dimensional features in the three branches to fully learn channel correlation.

The channel attention module[20] generates a description of each channel through a global pooling operation and calculates the channel attention weight through a shared multi-layer perceptron (MLP), it is shown in Figure 6. The specific steps are as follows:

$$F_{avg}=AvgPool(F)\in R^{1\times1\times C} \qquad (2)$$

$$F_{max}=MaxPool(F)\in R^{1\times1\times C} \qquad (3)$$

$$M_c(F)=\sigma(MLP(F_{avg})+MLP(F_{max}))\in R^{1\times1\times C} \qquad (4)$$

Where σ is the sigmoid activation function.

$$F'=M_C(F)\odot F \qquad (5)$$

Where ɯ represents element-by-element multiplication.

The spatial attention module generates spatial attention weights by averaging and max-pooling the channels of the feature map.

$$F'_{avg}=AvgPool(F')\in R^{H\times W\times 1} \qquad (6)$$

$$F'_{max}=MaxPool(F')\in R^{H\times W\times 1} \qquad (7)$$

$$M_S(F')=\sigma(Conv([F'_{avg},F'_{max}]))\in R^{H\times W\times 1} \qquad (8)$$

$$F'=M_S(F')\odot F' \qquad (9)$$

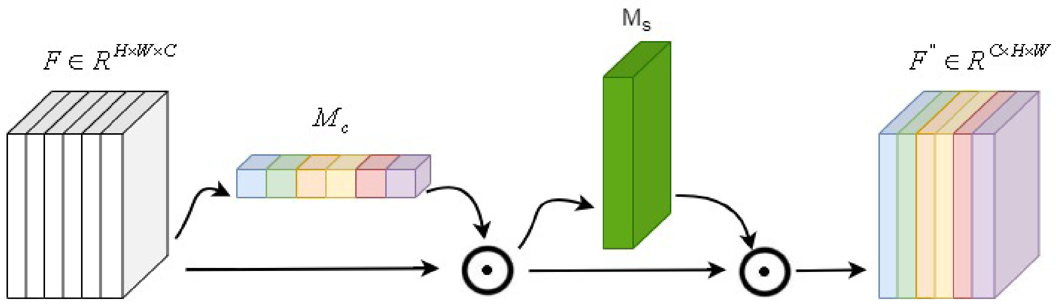

Figure 6 Attention mechanism module

C. Exit Flow

The proposed exit flow is exactly the same as that in original Xception model ,which consists of a classical convolution block and 1 residual separable convolution block.

Finally, we combine the fused features, and feed them to a fully connected layer for classification.

4 Experimentation and Results

4.1 Experimentation

4.1.1.Dataset

The FER2013 dataset[21] was employed to evaluate the performance of our proposed models.

Facial Expression Recognition 2013(FER2013) is a facial expression dataset released in the 2013 Kaggle competition "Challenges in Representation Learning: Facial Expression Recognition Challenge". It contains 35,887 facial images with expression categories annotated. The dataset is divided into three categories, including 28,709 training sets, 3,589 public test sets, and 3,589 private test sets. Each image is composed of a grayscale image with a fixed size of 48 ' 48. There are 7 expressions in total, corresponding to digital labels 0–6, including: 4953 angry, 547 disgust, 5121 fear, 8989 happy, 6077 sad, 4002 surprise, and 6198 neutral. This dataset does not directly give pictures, but saves the expression, picture data, and purpose data into csv files. When using it, the picture data value can be synthesized into pictures through code.The example of each expression in the FER2013 dataset is shown in Figure 7.

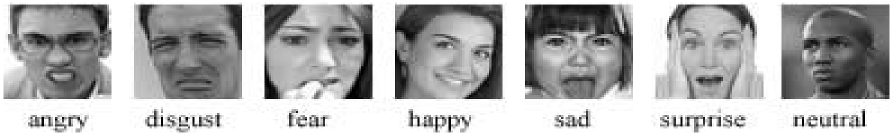

Figure 7 Seven samples of Fer2013 dataset

4.1.2Model Performance Evaluation Metrics

This article uses the most common evaluation indicators for facial expression recognition: accuracy, precision, recall and F1–score to evaluate the performance of the proposed model.

Accuracy reflects the overall correct classification of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

But in the case of class imbalance (i.e., some classes have far more samples than others), accuracy can be misleading.

Precision measures the proportion of correct predictions among all samples predicted to be positive.

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \qquad (11)$$

Recall measures the model´s ability to identify positive examples.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \qquad (12)$$

F1–score is the harmonic average of precision and recall, which comprehensively measures precision and recall. It is a single value that weighs these two metrics and is especially useful in cases of class imbalance.

$$F1 - score = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (13)$$

Where TP (True Positive) is the number of samples that are correctly predicted as positive examples. FN (False Negative) is the number of samples that are actually positive but are predicted to be negative. FP (False Positive) is the number of samples that are actually negative but are predicted to be negative The number of samples predicted as positive examples. TN (True Negative) is the number of samples correctly predicted as negative examples.

4.2Analysis of experimental results

The experimental system parameters are as follows:Ubuntu 18.04, 64-bit Linux system, Tesla T4 graphics card with 32 G memory, PyTorch 1.10, and CUDA 10.1.

4.2.1Validation Experiment results

We set the basic parameters:epoch = 50, learning rate = 0.001, batch_size = 64, weight_decay = 0.001, step_size = 20, the accuracy curve of the proposed model is shown as in Figure 8. When the model reaches 20 epochs, the training accuracy basically stops and approaches a stable value.

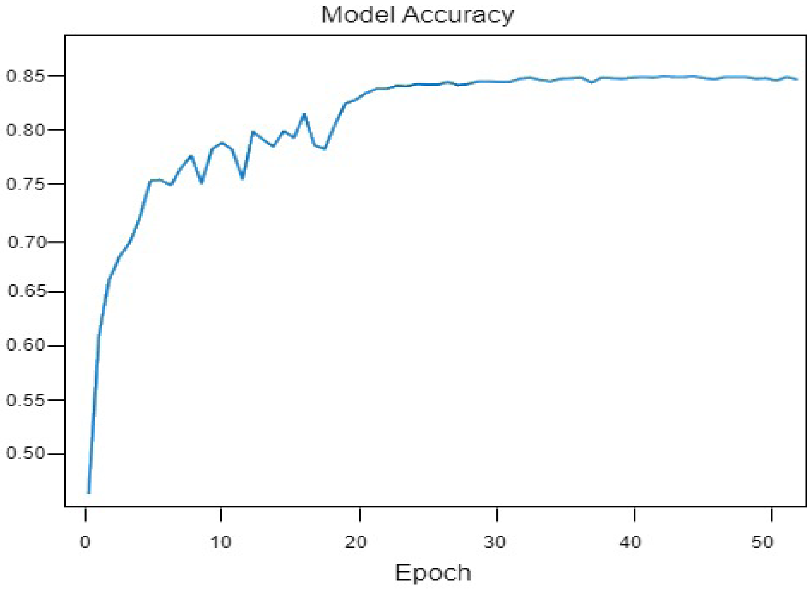

Figure 8 the accuracy curve of the proposed model on the training dataset

Table1 Performance comparison of seven lightweight CNN models

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score
The proposed model	84.45%	77.53%	82.91%	80.13%

Fig. 8 Confusion matric on FER2013 dataset

	angry	disgust	fear	happy	sad	surprise	neutral
angry	778	20	46	4	14	65	31
disgust	4	84	2	0	10	6	5
fear	23	28	883	9	22	36	23
happy	29	16	15	1614	9	29	62
sad	26	33	22	5	1124	19	18
surprise	21	17	33	10	21	715	14
neutral	54	67	43	23	35	147	864

4.2.2Comparative Experiment Results

To further demonstrate the advantages of our model, we compared the accuracy rate of six classic CNN models in Table 2.

Table 2 Performance comparison of seven classic CNN models

Model	Reference	Average accuracy(%)
VGG16	Ying Zhang et. al.[3]	71.31%
ResNet50	Mei Wang et. al.[22]	77.69%
Inception V1	Peng Zhang et.al.[23]	69.49%
Inception V2	Christian Szegedy [24] et.al.	76.60%
Xception	Qianqian Chen et.al.[25]	79.51%
The proposed model		84.45%

Table 1 shows that the improved Xception has the highest accuracy rate of 84.45% for the recognition of 6 facial expressions on the FER2013 dataset. The accuracy rate of VGG16 network is lower than the proposed model by 18.23%. The accuracy rate of ResNet50 network is lower than the proposed model by 11.85%.The accuracy rate of Inception V1 network is lower than the proposed model by 20.05%.The accuracy rate of Inception V2 network is lower than the proposed model by 12.94%.The accuracy rate of Xception network is lower than the proposed model by 10.03%.It can be seen that the multichannel convolutional neural network based on the fusion of attention mechanisms can significantly improve the accuracy of face recognition.

5 Conclusion

The Xception network reduces the number of parameters of the convolutional neural network and improves the network representation ability to a certain extent. However, the Xception network uses the same 3 ˙ 3 convolution kernel, and the image features extracted are relatively single. The single–channel attention module only acts on the channel information in the expression image, and has certain limitations in improving network performance. This model uses a convolutional attention module that focuses on useful feature information in both channels and space. It also improves Xception with multi–scale convolution kernels and multi–channel convolution.

REFERENCES:

[1] Voleti, S., et al., Stress Detection from Facial Expressions Using Transfer Learning Techniques, in 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT). 2024. p. 1–6.

[2] Porusniuc, G.C., et al., Convolutional Neural Networks Architectures for Facial Expression Recognition, in 2019 E–Health and Bioengineering Conference (EHB). 2019. p. 1–6.

[3] Zhang, Y., et al., Research on Facial Expression Recognition Algorithm Based on Deep Learning, in 2022 5th World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM). 2022. p. 1010–1013.

[4] Rawat, W. and Z. Wang, Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput, 2017. 29(9): p. 2352–2449.

[5] Xu, X., et al., A High–Precision Classification Method of Mammary Cancer Based on Improved DenseNet Driven by an Attention Mechanism. Comput Math Methods Med, 2022. 2022: p. 8585036.

[6] Kortli, Y., et al., Face Recognition Systems: A Survey. Sensors (Basel), 2020. 20(2).

[7] Xiao–Ying, G., et al., Multi–Granularity Feature Fusion Network for Dynamic Sequential Facial Expression Recognition, in 2023 China Automation Congress (CAC). 2023. p. 7064–7069.

[8] Luo, X., et al., Multi–Scale Local Feature Fusion Network for Facial Expression Recognition, in 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). 2022. p. 728–731.

[9] Li, W., et al., A Novel Multi–Feature Joint Learning Ensemble Framework for Multi–Label Facial Expression Recognition. IEEE

Access, 2021. 9: p. 119766–119777.

［10］ Lin, H., et al., Improved Xception with Dual Attention Mechanism and Feature Fusion for Face Forgery Detection, in 2022 4th International Conference on Data Intelligence and Security (ICDIS). 2022. p. 208–212.

［11］ He, Y., et al., Facial Expression Recognition Using Hierarchical Features With Three-Channel Convolutional Neural Network. IEEE Access, 2023. 11: p. 84785–84794.

［12］ Zhao, G., H. Yang, and M. Yu, Expression Recognition Method Based on a Lightweight Convolutional Neural Network. IEEE Access, 2020. 8: p. 38528–38537.

［13］ <Xception Deep Learning with Depthwise Separable Convolutions.pdf>.

［14］ Hardjadinata, H., R.S. Oetama, and I. Prasetiawan. Facial expression recognition using xception and densenet architecture. in 2021 6th International Conference on New Media Studies (CONMEDIA). 2021. IEEE.