Real-time facial emotion recognition using lightweight convolution neural network

Qi Li\textsuperscript{1,a}, Yun Qing Liu\textsuperscript{1,b\*, Yue Qi Peng\textsuperscript{1,c}, Cong Liu\textsuperscript{1,d}, Jun Shi\textsuperscript{1,e}, Fei Yan\textsuperscript{1,f}, Qiong Zhang\textsuperscript{1,g}

\textsuperscript{1}Changchun University of Science and Technology, Changchun, Jilin, China
\textsuperscript{a}email: liqicust@163.com, \textsuperscript{b\*}email: mzlyq@cust.edu.cn,
\textsuperscript{c}email: 736264660@qq.com, \textsuperscript{d}email: CongLiuCUST@163.com,
\textsuperscript{e}email: sj1607332877@163.com, \textsuperscript{f}email: yanf@cust.edu.cn,
\textsuperscript{g}email: 1119365209@qq.com

Abstract: In recent years, facial expression recognition has played an important role in the field of human-computer interaction, and the application of deep learning technology has enabled it to develop more rapidly. In this paper, we create a lightweight network model for real-time emotion classification to recognize student facial expressions. The system includes: using Haar cascade for face detection, combined with the idea of Xception to propose a lightweight CNN network model, using pre-activation in the residual block to optimize the model and reduce the impact of overfitting. The experimental results on the FER2013 database show that our model has a better expression classification effect than other state-of-the-art methods. Besides, our model uses fewer parameters, which reduces the complexity of network training. When real-time emotional recognition of students, the system can help teachers adjust their teaching methods according to the emotional state of students.

1. Introduction
Facial expression is the most direct and most effective emotion recognition model. With the development of artificial intelligence, research hotspots in the field of human-computer interaction have gradually increased. Emotion recognition has attracted people's attention in human-oriented design. In fact, some studies have shown the importance of machines to explain human emotions [1]. As a form of nonverbal communication, facial expressions can be understood faster than verbal communication. Due to the diversity and complexity of facial expressions, it includes several disciplines such as psychology and physiology and facial expression recognition. The developmental speed is slow and the categories are not completely independent of each other, which creates a lot of difficulty in recognizing facial expressions. Ekman and Friesen [2] proposed the Facial Action Coding System (FACS), which describes expressions through the combination of facial motion units so that more refined expression analysis can be obtained. FACS defined seven groups of facial expressions which are angry, disgust, scared, happy, sad, surprise, and neutral.

Emotions play an important role in students' learning and achievement. Emotions can also reflect students' attention in the classroom, learning motivation, and feedback on the acceptance of classroom content [3]. Teachers can adjust their teaching strategies and teaching materials according to the emotions conveyed by students' facial expressions to promote students' learning. In the past decade,
deep learning has dominated the field of artificial intelligence (AI), and convolutional neural networks have been widely used for image recognition and classification tasks. Several effective CNN architectures have been proposed at this stage, InceptionV3 [4] solves the problem of huge data volume and tedious network structure of traditional CNN by introducing sparsity and replacing fully connected layers with sparse connections and uses global average pool to reduce the number of parameters. XCEPTION [5] proposed that the cross-channel correlation and spatial correlation in the feature map of the convolutional neural network can be completely decoupled, and the Inception module in Inception v3 is replaced with a deep separable convolution [6], in the ImageNet data set and The best performance was produced on a large image dataset containing 350 million images and 17,000 categories.

Ayvaz et al. [7] developed a Facial Emotion Recognition System (FERS), which recognizes the emotional state and motivation of students in video conferencing e-learning. The system uses 4 machine learning algorithms (SVM, KNN, random forest, classification, and regression trees), and uses KNN and SVM algorithms to obtain the highest accuracy. S. Minaee [8] proposed to use an attention network in facial expression recognition research. The main idea is to add a space transformer module outside the typical convolution, pooling, and a fully connected layer of CNN. This module is used to transform the feature map to focus on the most relevant areas of the image, without the need for additional training or optimization modifications. Their approach CK + and FER2013 data sets achieved 98% and 70% precision. Nguyen et al. [9] proposed a multi-layer 18-layer CNN model similar to VGG. These models have not only high-level functions but also intermediate functions. On the Fer2013 data set, the accuracy of the ordinary CNN model reached 69.21%, and the accuracy of the proposed multi-level CNN model reached 73.03%. Chang et al. [10] used a CNN model based on ResNet to extract features from Fer2013 and CK+ datasets. The proposed complexity-aware classification algorithm (CPC) is applied with different classifiers (Softmax, LinearSVM, and RandomForest). The recognition accuracy rates of CNN + Softmax with CPC for Fer2013 and CK+ are 71.35% and 98.78%, respectively. The purpose of this article is to use a light-weight convolutional neural network to perform a more accurate real-time automatic analysis system for student facial expressions and realize emotion recognition in education.

2. Materials and Methods

In this section, this article will introduce the proposed system, which uses a deeply separable convolutional architecture to analyze the facial expressions of students. First, the system uses the Haar cascade classifier to detect face information from the input image, and crops and normalizes the detected face information to a size of 48×48. Then, these facial images are used as input to CNN. Finally, the output is the facial expression recognition result (happy, sad, anger, fear, surprise, disgust and normal). Figure 1 shows the structure of the method proposed in this paper.

![Figure 1. Facial expression recognition system structure diagram](image-url)
In this paper, build a new model structure CNN mainly consider three factors: speed training, recognition accuracy and memory consumption. The network is mainly composed of seven convolutional layers, a maximum pooling layer and a global average pool. The algorithm idea and model construction process of this network are shown in Figure 2:

First, the image of the input layer adopts a preprocessed grayscale image of a face with a size of $48 \times 48$, selects 8 convolution kernels with a size of $3 \times 3$, and performs convolution processing on the input image to obtain a local expression feature map. And use the ReLU function for nonlinear activation.

Secondly, when convolving the upper layer to output the feature map, two convolution kernels with different sizes of $3 \times 3$ and $5 \times 5$ are used to achieve a more detailed extraction of facial expression images of different scales.

Third, this network uses 4 residual depth separable convolutions to reduce the cost and number of parameters of convolution. The difference between standard convolution and depth separable convolution is shown in Figure 3. Depth separable convolution is completed in two stages: Depthwise Convolution and Pointwise Convolution. In the first stage, in the depth operation, convolution is applied to a single channel at a time, instead of convolving all M channels like standard CNN. Therefore, the size of the convolution kernel here is $D_k \times D_k \times 1$. Assuming that there are channels in the input data, one such filter is needed, and the output size is $D_p \times D_p \times M$. Then the total number of operation multiplications of deep convolution is as equation (1):

$$DC = M \times D_p^2 \times D_k^2$$  \hspace{1cm} (1)

In the pointwise convolution operation, a $1 \times 1$ convolution operation is applied on $M$ channels. Therefore, the size of the convolution kernel for this operation will be $1 \times 1 \times M$. Assuming we use $N$ convolution kernels, the output size is $D_p \times D_p \times N$. Then the total number of multiplications for pointwise convolution operations is as equation (2).

$$PC = M \times D_p^2 \times N$$ \hspace{1cm} (2)

For the same input image, the number of multiplications of the depth separable convolution is the sum of the number of multiplications in the two stages. By comparing the calculation amount of the depth separable convolution with the calculation amount of the standard convolution, as shown in equation (3), the multiplication and the available total number of training parameters are reduced:

$$\frac{[\text{depthwise}] \times [\text{separable}] \times [\text{convolution}]}{[\text{standard}] \times [\text{convolution}]} = \frac{M \times D_p^2 \times (D_k^2 + N)}{N \times D_p^2 \times D_k^2 \times M} = \frac{1}{N} + \frac{1}{D_k^2}$$ \hspace{1cm} (3)
Fourth, deep neural networks often suffer from the problem of vanishing gradients, that is, the gradient signal from the error function will drop exponentially when it propagates back to an earlier layer. In this paper, the network uses the gradient signal in the residual module to achieve the output of one layer as input to the other layer of the architecture through "skip connections" or "shortcut path", which is activated by adjustment. The position of the function, the activation function is placed before the final output, that is, the input value is normalized after the input is obtained, and then activated, and then multiplied by the weight layer, and finally output. Make the parameters of the network easier to learn. Figure 4 shows an example of the Pre-activate the residual block. In this way, the learned feature is the difference between the original feature map and the desired feature map. The residual block does not increase the complexity of the model, nor does it increase the number of trainable parameters.

Figure 3. (a) Standard Convolution (b) Depth Separable Convolution

3. Results & Discussion
The work in this paper is implemented using Python3 and Keras2 to run on Tensor-Flow with GPU (Nvidia Quadro P5000) support.

This article selects the FER2013 data set provided by the 2013 Kaggle facial expression analysis contest. The data set contains 28709 training samples, 3859 validation data sets and 3859 test samples, a total of 35887 images containing seven categories of angry, disgust, fear, happiness, sadness, surprise and normal. The image resolution is 48×48. An example of the FER2013 data set is shown in Figure 5. The artificial accuracy of this database is 65% ± 5%. Since the FER2013 data set is more complete and more in line with real-life scenarios, this article mainly selects the FER2013 training and testing model. To prevent the network from overfitting too quickly, the model in this paper transforms the data set image, such as flipping, rotating, cutting and other data enhancement operations.
The experiments in this article are conducted on FER2013 database samples for training and testing. The network parameters are preset by random initialization, and then standard backpropagation is used to optimize these parameters. The average of the results of more than 100 cycles is used as the test result. The accuracy of this model on the training set and validation set is shown in Figure 6, and the loss is shown in Figure 7. It can be seen from the figure that the model gradually stabilizes and reaches the maximum value after the 100th epoch of training.

The confusion matrix of the model in this paper is shown in Figure 8. We can observe that predicting "happy" expressions is more accurate, reaching an accuracy of 91%. Since the number of smiley images in the training set is larger, accounting for 25% of the data set, the highest accuracy appears. The disgust data in the data set only accounts for 1.5%, the classification accuracy is still slightly higher than the overall accuracy, reflecting the effectiveness of the model. At the same time, there are several common misclassifications. For example, an image with an expression of "sad" is misidentified as a "scared" expression, and a predicted "anger" is misidentified as an expression of "disgust". Overall, the model architecture proposed in this paper achieves 72.4% accuracy on the FER2013 data set.
In this paper, four young master students participated in the system accuracy experiment. Each student selected a single frame image within 10 seconds to predict emotions. The predicted classification results are shown in Table 1. The emotion labels in the images are represented by blue boxes. Indicates that the percentage value represents the probability of emotion. The experimental results show that the real-time classification results of the system are more accurate.

Table 1. Students’ facial emotion recognition results

| Emotion | Real-time Emotion Recognition |
|---------|-------------------------------|
|         | Angry | Disgust | Scared | Happy | Sad | Surprised | Neural |
|         |       |         |        |       |    |           |        |
|         | 52.51%| 3.98%   | 13.91% | 1.10% | 9.53% | 1.99%     | 16.98% |
|         | 25.75%| 57.73%  | 10.23% | 0.02% | 5.84% | 0.06%     | 0.36%  |
|         | 7.76% | 0.36%   | 60.72% | 1.89% | 1.79% | 26.84%    | 0.64%  |
|         | 0.00% | 0.00%   | 0.02%  | 99.38%| 0.01% | 0.07%     | 0.93%  |
|         | 27.35%| 1.09%   | 4.00%  | 4.57% | 37.00%| 1.10%     | 24.90% |
7.41% 0.93% 16.92% 16.24% 8.72% **48.02%** 1.76%

1.95% 0.03% 5.90% 5.79% 2.87% 3.42% **80.04%**

4. Conclusions
This paper proposes a convolutional neural network model for real-time recognition of student facial emotions. The model uses deep separable convolution to reduce the number of training parameters in the convolution operation. The model only contains about 58,000 parameters, which reduces the complexity of the model. The accuracy of the model on the FER2013 database is 72.4%. The facial expression recognition system in this paper detects the facial image and divides it into 7 basic expression categories. This system can assist teachers to make correct teaching adjustments to students' understanding in the classroom. The next research will further improve the model architecture to improve the classification performance, and use 3D expression images for convolutional neural network training to obtain more accurate emotion recognition.

Acknowledgments
This work was supported by the Science and Technology Department Project of Jilin Province (under grants No. 20190303034SF and No. 20180201064GX).

References
[1] Kaiser, R., Oertel, K., (2006) Emotions in HCI: An Affective e-Learning System. In: Proceedings of the HCSNet Workshop on Use of Vision in Human-computer Interaction, Darlinghurst, pp. 105–106

[2] Saurav S, Singh S, Saini R, et al. (2016) Hardware accelerator for facial expression classification using linear svm. Springer, 425(01):39-50.

[3] Lasri, I. Solh, A.R., Belkacemi, M. E., (2019) Facial Emotion Recognition of Students using Convolutional Neural Network. In: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), Xian

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 2818-2826

[5] Francois, C., (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, pp. 1800-1807

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., & Weyand, T., et al. (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. Computer Vision and Pattern Recognition (cs.CV), arXiv:1704.04861 [cs.CV]

[7] Ayvaz, U., Hüseyin Gürüler, & Devrim, M. O., (2017) Use of facial emotion recognition in e-learning systems. Information Technologies and Learning Tools, 60(4): 2076-8184.

[8] Minaee, S., & Abdolrashidi, A., (2019) Deep-emotion: facial expression recognition using attentional convolutional network.

[9] Chang, Tianyuan & Wen, Guihua & Hu, Yang & Ma, Jiajiong. (2018) Facial Expression Recognition Based on Complexity Perceotion Classification Algorithm.

[10] Nguyen, H. D., Yeom, S., Lee, G. S., Yang, H. J., & Kim, S. H., (2018) Facial emotion recognition using an ensemble of multi-level convolutional neural networks. International Journal of Pattern Recognition and Artificial Intelligence (1).