Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

YOLO-based Face Emotion Recognition System for Teaching Scenes

Yuxuan Wang*

School of Electronic and Information Engineering, Anhui University, Hefei, China

* Corresponding Author: P02114039@stu.ahu.edu.cn

Abstract. The main way to get emotional data has been through how they looked, which are able to interact a variety of feelings like joy, grief, fear, rage, disgust, and surprise. These expressive emotions are vital to communication due to their enable people to better understand one another's thoughts and desires. Therefore, we thought that we could utilize technologies such as computer vision to obtain images of students' expressions in the classroom through cameras and other related devices. In order to create a database, the camera is used to build an image acquisition system. One image is preprocessed, the YOLOv5s algorithm is used to extract the features of the face, and the outcome of correct recognition is determined by comparing the feature vectors that were obtained. The expressions are then categorized using the extracted feature vectors based on a multilayer perceptual network (MLP). Lastly, the computer summarizes the identifying data and uses a WeChat applet to deliver it to the teacher.Teachers can more effectively comprehend the educational needs of their pupils surroundings and modify their approaches to instruction with the help of this WeChat applet. This will help improve the quality of online education.

Keywords: YOLOv5s; Multilayer perceptual; Learning state assessment; Database.

1. Introduction

The primary manner in which to get emotional information is through facial expressions, which can communicate a variety of feelings like joy, sorrow, fear, rage, disgust, and surprise. These expressive emotions help people to better understand each other's emotions and intentions and play an important role in the communication process. Studies have shown that facial expressions contain a wealth of social information, reflecting the truest emotional state of the expresser, and are one of the most representative signals for human beings to express their emotional states and intentions [1]. As a result, emoji information has also been widely used in fields such as user experience evaluation and psychology to help gain a deeper understanding of the nature of human behavior and thinking [2]. The advent of algorithms that use deep learning and the availability of massive amounts face expression databases resulted in notable advancements during the area of face expression identification in recent years [3]. Researchers have begun to study facial expression recognition technology, which analyzes the information conveyed by facial expressions on a person's face in order to better understand people's behavior or provide a targeted response. Various types of human facial expressions are shown in figure 1.

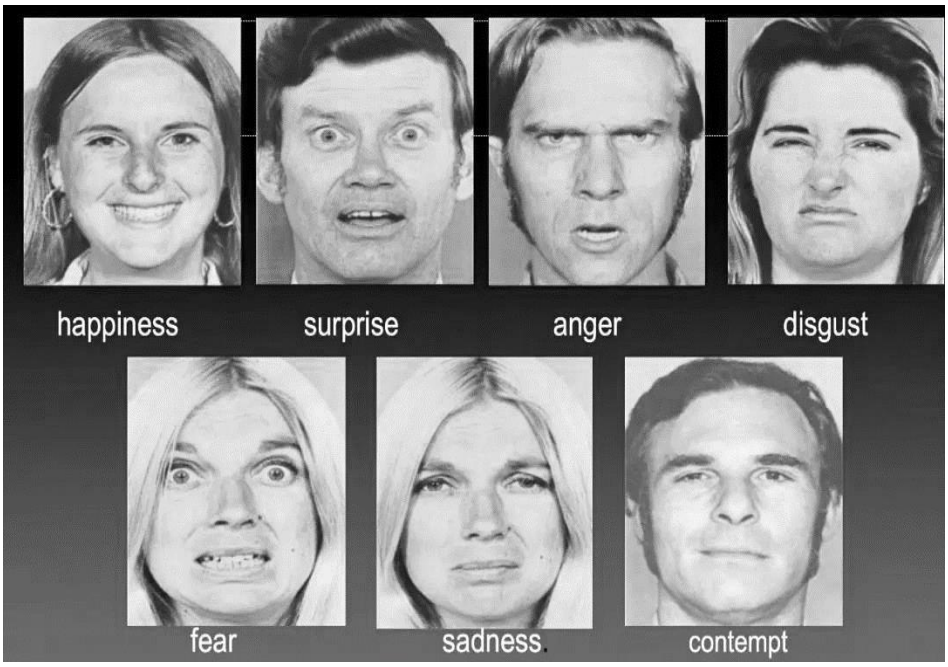

Figure 1. Various types of human facial expressions

The quick advancement of technological innovations has encouraged the variety of teaching approaches as well as altered the distribution of education, making it more practical, effective, and thorough. In traditional classroom education, teachers can assess students' mastery of course content by observing their emotional state. If the majority of students are in a positive mood and actively participating, then it means that they have largely grasped the content of the class and the teacher can continue to follow the lesson plan. On the contrary, if most of the students are in a negative mood and distracted, then it means that they are not keeping up with the teacher's pace of instruction and are experiencing learning disengagement. At this point, teachers need to adjust their teaching methods in time to improve the effectiveness of teaching. With the continuous maturity of artificial intelligence and other technologies, “artificial intelligence + education” provides us with a new direction to solve the problem of student learning disengagement in online education [4]. For the assessment of students' learning status in online education environments, we can utilize technologies such as computer vision to obtain images of students' expressions in the classroom through cameras and other related devices [5]. Then, a face expression recognition model is used to recognize students' emotional states and map them into learning states for real-time feedback to the teacher. This can help teachers and students to determine the learning status of students in the classroom, thus improving the teaching effectiveness and quality of online education. By monitoring and analyzing students' emotional state, teachers can improve their effectiveness in teaching by promptly modifying their techniques based on a more precise comprehension of the educational circumstances of their pupils. This will help improve the quality of online education.

2. Methodology

2.1. Technological path

This project is oriented to online education, artificial intelligence technology and other research hotspots, in response to the teaching needs of the network platform, supervise the learning status of each student, and carry out the establishment of a learning status assessment system based on the emotion recognition of the face and the emotion classification of the learning status. The project intends to adopt the research method of experimental research, combining theoretical analysis and experimental verification. And the technical route is shown in figure 2.

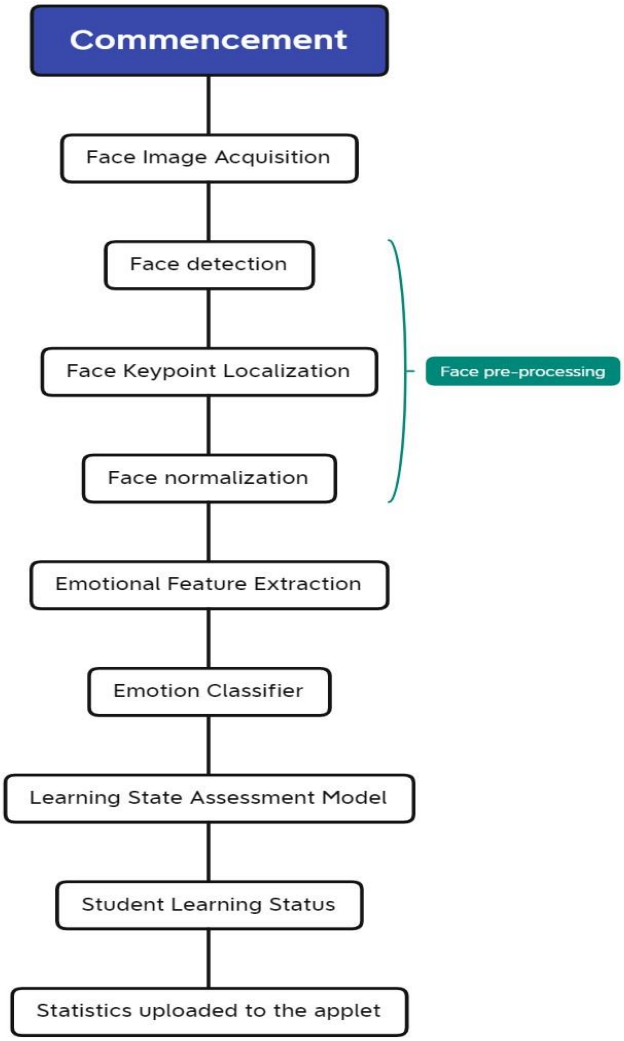

Figure 2. Flowchart of technical route

2.2. Research program

2.2.1. Face preprocessing based on image acquisition system

Since there is no publicly available dataset for classroom mood recognition analysis of students, in order to accurately identify the change of each student's mood state, this design records videos of students in real teaching and learning environments through an image capture system and intercepts them as images for data acquisition. Since the classroom environment is relatively homogeneous and the change of face location is small, this paper selects a time interval of 10s, i.e., extracting a picture from the video every 10s as a sample for face detection.

The original image captured through the camera is subject to external conditions (e.g., light), and directly used for face recognition will produce poor results, so it cannot be directly applied to face recognition, so the original image needs to be further preprocessed before face detection.

① Gray-scale transformation： To enhance the image's clarity, each gray color in the source photograph was subjected to linear, segmented linear, and nonlinear transformations.

② Normalization： Due to the influence of image acquisition equipment and other external factors, the captured images are not completely consistent, so it is necessary to use normalization algorithms to unify the images by cropping, positioning, rotating, and scaling.

③ Denoise：In order to increase the efficiency of image recognition to reduce the other disturbances in the image, so the image needs to be processed by denoising operation.

In this project, the proposed boundary-aware face alignment network is used to extract the key points of the face as shown in figure 3.

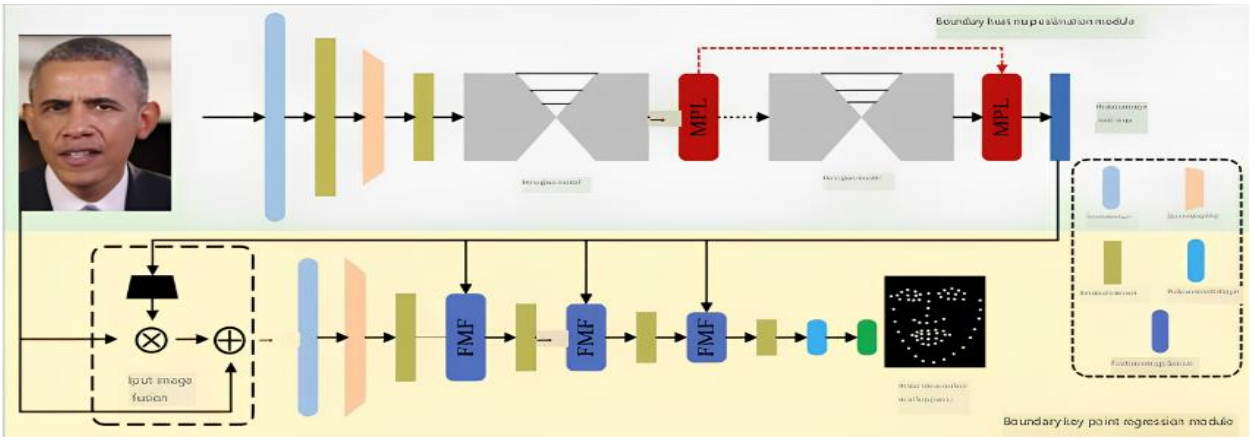

Figure 3. Boundary-aware face alignment network

The network consists of two main templates: a boundary heat map estimation template and a boundary key point regression template. The Boundary Heat Map Estimation module generates the boundary heat map as a geometric structure of the face. In order to forecast keypoint supervises, the Boundary Key point Regression module uses data from the boundary in a several phases method.

Given a face image I, its true annotation is represented by L given the keypoints as $S = \{s_l\}_{l=1}^{L}$. The K subsets $S_i \subset S$ represent the keypoints belonging to the K-boundary, respectively. Packed lines of boundary are obtained by applying interpolation to S_i for each boundary. By changing the other stores to 0 and only the points that are on the edge of the line to 1, one can create an integer barrier image B_i that is the identical size as I. Perform a distance transformation in line with each B_i to obtain the separation mapping D_i. The chart of distance is converted to a border of reality heat map M_i using a Gaussian expression for the standard deviation σ.

$$M_i = \begin{cases} \exp\left(-\frac{D_i(x,y)}{2\sigma^2}\right), if D_i(x,y) < 3\sigma \\ 0, otherwise \end{cases} \tag{1}$$

The boundary heat map M is fused with the input photograph I. The definition of the merged input H is:

$$H = I \oplus (M_1 \otimes I) \oplus ... \oplus (M_T \oplus I) \tag{2}$$

where $\oplus$ represents the channel connection and $\otimes$ the dot product operation of the elements. The boundary heat map M is fused with the feature diagram F. The definition of the hybrid map of characteristics H is:

$$H = F \oplus \left(F \otimes T(M \oplus F)\right) \tag{3}$$

For the generated boundary heat map $\hat{M}$, set the coordinates of its corresponding generated keypoints to $\hat{s}$ and denote the distance to ground truth matrix image as $Dist$. The method of discrimination D fundamental truth value d_{fake} is defined as follows:

$$d_{fake}(\widehat{M},\ \hat{s}) = \begin{cases} 0, & Pr_{s\in\hat{s}}(Dist(s) < \vartheta) < \delta \\ 1, & otherwise \end{cases} \tag{4}$$

where ϑ is the distance threshold from the boundary of the truth value of the surface, where δ is the probability cutoff. By combining a boundary valid row discriminator D with a boundary heat map estimator G, the idea of adversary instruction is presented. The loss of D can be expressed as:

$$L_D = -\left(E[log\, D(M)] + E\left[log\left(1 - \left|D\left(G(I)\right) - d_{fake}\right|\right)\right]\right) \tag{5}$$

where M is the boundary heat map of the ground reality. While predicting the generated boundary heat map based on d_{fake}, the discriminator gains the ability to forecast the ground truth boundary heat map overall. An expression for the adversarial loss using the effectiveness discriminator is as follows:

$$L_A = E\left[log\left(1 - D\left(G(I)\right)\right)\right] \tag{6}$$

Through the positioning of facial feature points, the key areas of the face (e.g., eyebrows, eyes, nose, mouth, etc.) are accurately located, and then the face is aligned to the predefined template according to the position information of the feature points, so that the normalized face has a uniform size and each part of the face has a corresponding correspondence.

2.2.2. Feature extraction based on YOLOv5s algorithm

YOLOv5s is a target detection algorithm based on the Anchor detection method, its main idea is to divide the whole image into a number of grids, each grid predicts the category and positional information of the objects in the grid, and then according to the IoU value between the predicted frame and the real frame for the filtering of target frames, and the final output of the predicted frames of the category and positional information [6].

The feature extraction network of the YOLOv5s model is divided into two parts: the backbone network and the neck. In which the backbone network part uses the common convolutional structure CBS module with residual structure C3 module and adds the fast spatial pyramid pooling (SPPF) at the end. See Figures 4, 5, and 6 for specific structures.

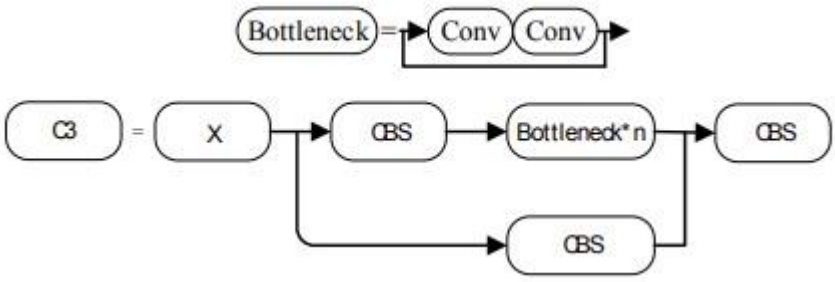

Figure 4. Diagram of the CBS module

Figure 5. Diagram of the C3 Module

Figure 6. SPPF module

One of them, the CBS module, is composed of a 2D convolution, batch normalization, and activation function; The C3 module draws on the cross-stage local network structure to further enhance the feature learning capability of the network; The SPPF module, on the other hand, effectively avoids image distortion caused by cropping and scaling operations on the image area.

The Neck structure, on the other hand, combines a Feature Pyramid Network (FPN) with a Path Aggregation Network (PANet), in which the FPN passes semantic information from top to bottom of the higher-level feature maps, which are smaller in resolution and carry high-level semantic feature information, thus enhancing semantic representation at multiple detection scales; PAN, on the other hand, conveys localization information from the bottom up by using low-level feature maps with larger resolution and carrying localization information to compensate for and enhance the localization capability at each detection scale.

The YOLOv5s model uses multi-scale detection to detect targets from three different scales, the three different detection scales come from three different outputs in the neck structure, when the size of the input image is (640×640), the size of the feature maps of three different depths are (80×80), (40×40), and (20×20), the bigger the feature maps the more likely it is to be able to detect the small scale targets, and vice versa. smaller feature maps are more likely to detect large targets. The YOLOv5s model overall structure diagram is shown in figure 7.

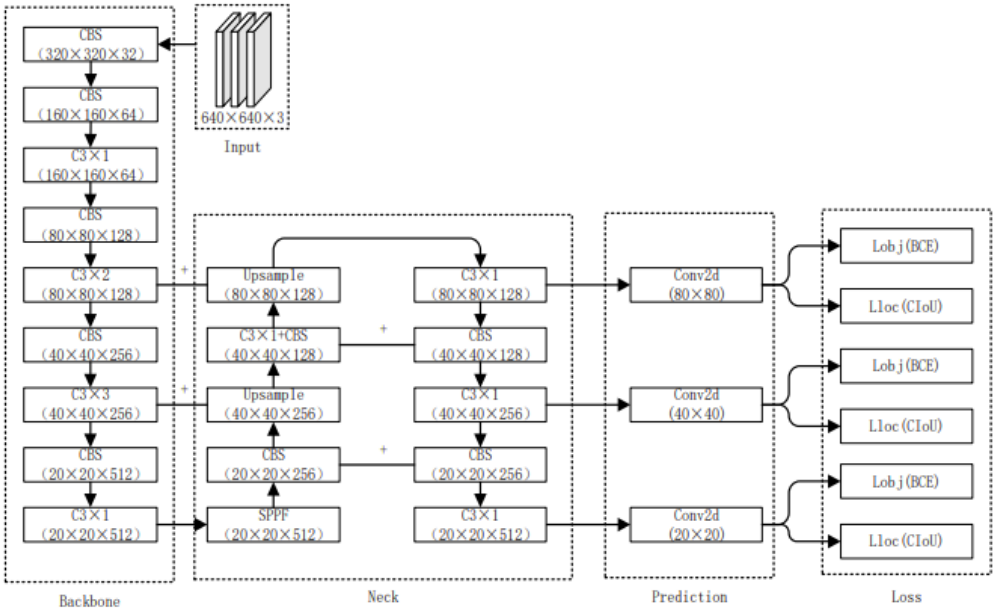

Figure 7. YOLOv5s model overall structure diagram

2.2.3. Categorization of expressions using the MLP emotion classifier

A multilayer perceptron is an artificial neural network based on a feed-forward neural network consisting of an input layer, a hidden layer, and an output layer with links between two-by-two neurons in adjacent layers [7]. The neural network architecture for MLP is shown in figure 8.

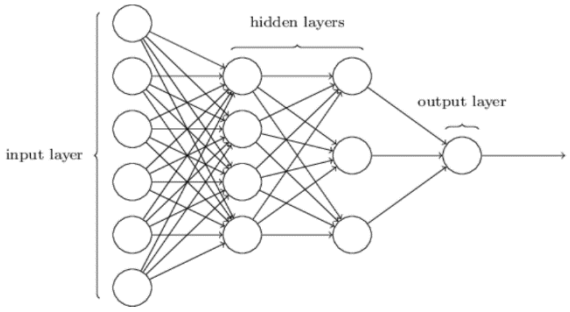

Figure 8. Neural Network Architecture for MLP

The dataset's feature vectors are obtained via the MLP's input layer, and each feature vector is then multiplied by a bias and its matching weight to create the hidden layer neurons; The result of all neurons of the last hidden layer multiplied by the weights and summed with the bias is the input of the output layer; The output layer maps its received inputs to get the mesh through an activation function and finally outputs the result. The commonly used activation functions are shown in figure 9.

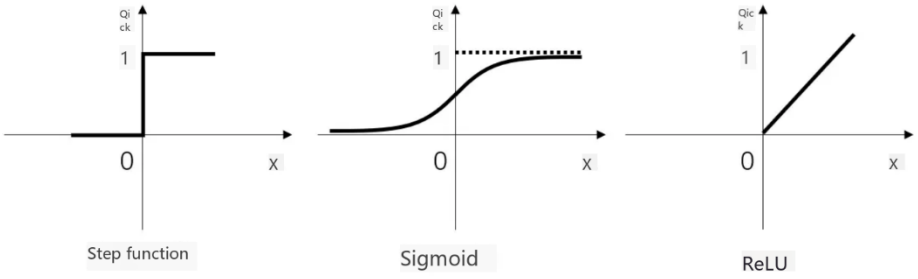

Figure 9. Commonly used activation functions

This system utilizes the MLP emotion classifier to determine whether the extracted feature vector is the defined emotion based on the similarity comparison with the known emotion features. After examining a range of different network topology parameters, the number of neurons in the input, hidden and output layers were proposed to be designed. One of the output layers is 6 neurons corresponding to the 6 basic emotional states of happiness, surprise, fear, sadness, disgust, and anger. The possibility of falling into a local optimum is reduced by employing a sigmoid function through the MLP and utilizing an inverse BP network for training. The specific operational flow of the MLP is as follows: Suppose the input layer is the x module, the hidden layer is the h module and the output layer is the y module as shown in Figure 10.

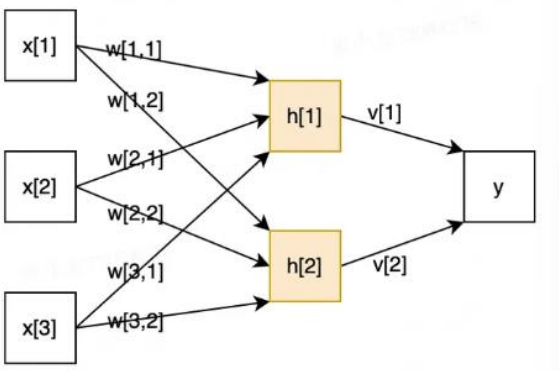

Figure 10. Network structure diagram of MLP

Then y and h are related to x:

$$h[1] = f(\omega[1,1]x[1] + \omega[2,1]x[2] + \omega[3,1]x[3] + b[0]) \quad (7)$$

$$h[2] = f(\omega[1,2]x[1] + \omega[2,2]x[2] + \omega[3,2]x[3] + b[1]) \tag{8}$$

$$y = v[1]h[1] + v[2]h[2] + s \tag{9}$$

Here $f(\cdot)$ employs a sigmoid function to map the output to between 0 and 1 with the expression [8]:

$$y = \frac{1}{1+e^{-x}} \tag{10}$$

Next the output values are cached and the BP is combined with optimization methods such as gradient descent to train the artificial neural network to achieve the effect of minimizing the prediction error of the MLP with respect to the sample. Let the loss for any sample be $Loss$, which uses the variance loss function, then the $Loss$ expression is:

$$Loss = [y - (v_1 f(\omega_{11}x_1 + \omega_{21}x_2 + \omega_{31}x_3 + b_0) + v_2 f(\omega_{12}x_1 + \omega_{22}x_2 + \omega_{32}x_3 + b_1) + s)]^2 \tag{11}$$

If there are a total of N samples in the training set, sum the error of each sample, i.e., $min \sum_{i=1}^{N} Loss_i$, in the above equation, the variables to be optimized are v, ω, b and s. In order to minimize the error sum, the gradient is solved by gradient class algorithm and the gradient is computed using the most rapid descent method in the following. As an example, the matrix form of the expression for the gradient of $Loss$ against ω is:

$$\frac{\partial h}{\partial \omega^T} = f' \cdot x \tag{12}$$

$$\frac{\partial \hat{y}}{\partial h} = v^T \tag{13}$$

$$\frac{\partial Loss}{\partial \hat{y}} = y - \hat{y} \tag{14}$$

This is obtained by the chain rule for derivation:

$$\frac{\partial Loss}{\partial \omega^T} = \frac{\partial Loss}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial h} \cdot \frac{\partial h}{\partial \omega^7} = (y - \hat{y}) \cdot v^T \cdot f' \cdot x \tag{15}$$

This results in a sample gradient value. To minimize the amount of computation required to achieve precise gradient values, this system is designed to select a part of the samples each time and calculate the gradient value as the final gradient value, so as to achieve the requirement of taking into account the efficiency of the calculation and the accuracy of the value by adjusting the number of samples. After obtaining the gradient, the most rapid descent method is performed:

$$\mathrm{d}_k = \frac{\partial Loss}{\partial \theta} = \nabla(\theta) \tag{16}$$

where θ denotes the optimization variable and d_k is the direction of the kth iteration. The loop proceeds until the specified number of iterations is reached, yielding the optimal value of the variable.

2.2.4. Learning state assessment model based on learning state space and face detection algorithm

In order to further make instructional improvements to the various emotional states, emotions were categorized into three categories based on the learning state space: positively contributing to learning, negatively contributing, and having an unclear impact [9]. The categorization of classroom students' learning states was achieved through a simple mapping of emotional states, i.e., sad, angry, and disgusted emotions corresponded to negative states, indicating that students' attention in class was not in the classroom or they were disgusted by the content of the lecture; happiness and surprise correspond to positive states, indicating that students are in a relatively high mood to learn and pay attention; and fear corresponds to the state of not being obvious.

The face detection algorithm characteristics are utilized to examine high-level human facial features (including the eyes, nose, and corners of the mouth on both sides). where faces with less than two features detected are considered to be low, i.e., completely unfocused level; identified two or three characteristics as being aware, looking up, but not looking directly at the board/teacher; identified four or more characteristics considered attentive and looking directly at the board/teacher. Meanwhile, emotion recognition only classifies emotions for faces that are able to recognize more than four features, and faces that do not recognize more than four features are classified as unrecognized.

This design simultaneously identifies a positive state with four or more features identified as excellent learning status; identifies a positive state or identifies four or more features identified as good learning status; identifies a negative state or detects fewer than two features, and simultaneously identifies a negative state with fewer than two features identified as poor learning status; unrecognized is judged as absent; and the rest of the cases are judged as average learning status. The rest of the cases were judged as fair learning status.

This design applet has a search page and a message page. Teachers can click on the search interface to search for the course they are teaching, and view the learning status of students of different levels in the three options of “Fluctuation of each student's learning status”, “Students with poor learning status (marked in red)”, “Active and serious students”, and “Students with low learning status (marked in red)”. “Teachers can check the learning status of students of different levels in the lesson by clicking on the search screen. Once you have accessed any page, you can view the individual student's learning for the session by clicking on the individual student's avatar. It is also possible to look up a student's name directly on the search page as a way to see how a particular student's learning status has changed from session to session. The initial interface of the applet is shown in figure 11.

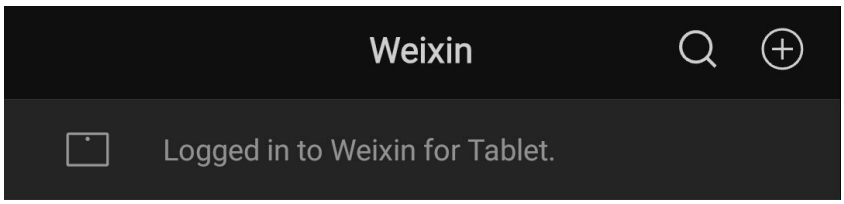

Figure 11. Small program search bar effect

In the message page, the system sends a list of students who have failed to meet the classroom learning status three or more times during the semester to the teacher's cell phone to remind him/her of the special attention. By responding to teachers about the fluctuation of each student's learning status in each lesson, teachers are able to fully understand the differences in the learning status of different students so that they can adopt different teaching methods, which greatly improves the quality of teaching.

3. Experiments

3.1. Experimental Setup

My model is optimized using a cross-entropy loss function and a learning rate scheduler. In deep learning, the cross-entropy loss function is used to measure the difference between the predicted distribution of a model and the distribution of real data as a way to evaluate the predictive performance of a model. In comparing the cross-entropy loss function with the variance loss function used in the initial stage, it is found that the variance loss function is more cumbersome and complex when activated by utilizing some saturated activation functions and can lead to problems such as unstable update magnitude. In contrast, a neural network model's learning rate can be automatically modified by the learning rate scheduler in accordance with the training process, which enhances the model's convergence and effectiveness. This model utilizes the torch.optim.Ir_scheduler module in PyTorch to implement the learning rate scheduler. Therefore, optimization with cross-entropy loss function and learning rate scheduler makes the model more accurate. And when using the dataset for training, I used a weighted sampling strategy with the weights set to the proportion of samples of different emotions. I chose macro-F1 and weighted-F1 scores as the main assessment metrics. Due to the unbalanced dataset, I also selected the arithmetic mean (macro-F1) and the weighted mean (weighted-F1) to measure the performance of each category. When the macro F1 performance does not improve within 50 epochs, training is stopped and the best results are reported [10].

3.2. Experimental Results

Table 1 displays the experimental findings derived from the LOSS, ACCURACY, Macro_F1, weighted_F1, the training results are relatively good, and it is more accurate in recognizing the faces in the dataset. The results of the arithmetic mean (macroF1) and the weighted mean (weighted-F1) are quite good, which proves that the performance of our overall method is more reliable and effective.

Table 1. Training results

Epoch	Loss	Accuracy	Macro_f1	Weighted-f1
128	0.0037	0.9993	0.9989	0.9992

3.3. Further Analysis

Based on the advantages of multimodal model with rich information fusion and wide application areas, I decided to optimize the face feature extraction and emotion classification based on Yolov5s algorithm by absorbing the advantages of MemoR model. The main areas include the following:

The Self-Attention System: A network configuration known as the Self-Attention System is primarily used in scenarios in which the network receives a large number of uncertainly-sized vectors as input. Examples of these scenarios include machine translation (the so-called sequence-to-sequence problem, in which the machine determines how many tags to use), lexical annotation (also known as Pos tagging, in which one vector corresponds to one tag), semantic analysis (in which multiple vectors correspond to one tag), and other word processing issues. With the use of a self-attention mechanism, this model more effectively represents the dependencies among various encoder components.

Multi-Layer Perceptron (MLP): Keeping in line with the initial conception, MLP, as a common deep learning model, accomplishes the operations of extracting features of input data to abstracting and classifying the features by interconnecting neurons at different levels. In this model MLP is used for nonlinear transformation of features.

Fusion Layer: A fusion layer is a technique that combines multiple operations or layers into a single operation or layer. This model utilizes Fused Convolution Kernel in PyTorch to enhance the performance of the simulation, the Fused Convolution Kernel is implemented by using

torch.nn.functional.con2d function in PyTorch. Partial screenshot of the model utilizing the fused convolution kernel is shown in figure 12.

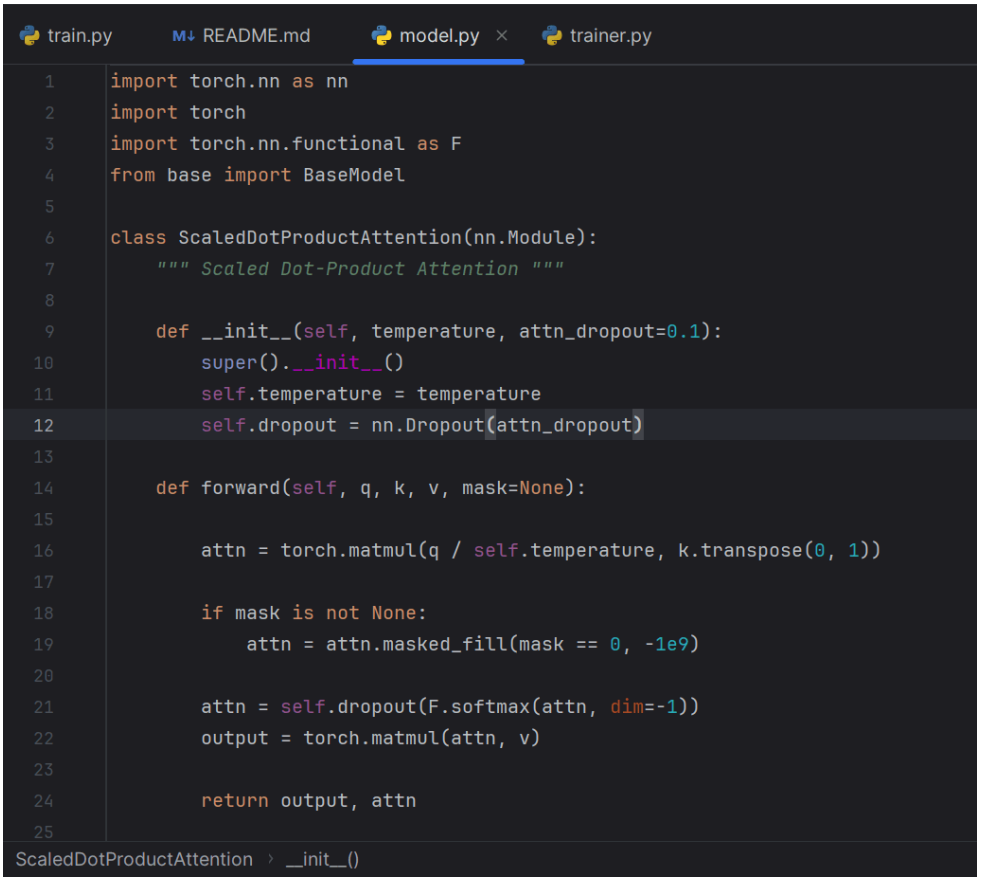

Figure 12. Partial screenshot of the model utilizing the fused convolution kernel

4. Conclusion

The past few decades have seen a remarkable advancement in artificial intelligence hardware and technique. researchers have begun to study Facial Expression Recognition (FER), which analyzes the information embedded in facial expressions of a person's face in order to better understand people's behaviors or give targeted responses. Because of this, I am quite interested in face recognition technology and made the above one system research. The face emotion recognition system for teaching scenarios constructed based on Yolo can be used in today's online education environment to solve the problem of low teaching effect of online education to a certain extent. This neural network-based automatic expression recognition system. It can recognize students' emotions in real time and feedback the results to teachers, helping them to better understand students' learning status. The system can realize the recognition of multiple expressions with high accuracy and practicality. This paper hope that the system can more accurately recognize the seven basic facial expressions as well as neutral expressions and help teachers better understand students' emotional and learning states for more refined teaching.

References

[1] Shan L, Weihong D. Research progress in deep face expression recognition. Chinese Journal of Image Graphics, 2020, 25(11): 2306-2320.

[2] Qiang H, Kai W. Analysis and prediction of user experience of airline official website based on “eye movement + facial expression. Science, Technology and Engineering, 2022, 22(20): 8739-8747.

[3] Chenqi L. Research and application of face expression recognition algorithm based on tensor representation and deep learning. University of Electronic Science and Technology, 2020.

[4] Lin Z. Exploration of changes and paths in the field of education based on artificial intelligence. Continuing Education Research, 2024, (01): 38-41.

[5] Lin W, Menglin L. Analysis of student concentration in online classroom based on face expression recognition. Computer System Applications, 2023, 32(02): 55-62.

[6] Sun X, Zhang X, Xia Z, et al. Advances in Artificial Intelligence and Security. Springer International Publishing, 2021.

[7] Yang P, Hu J, Hu B, et al. Estimating soil organic matter content in desert areas using in situ hyperspectral data and feature variable selection algorithms in southern Xinjiang, China. Remote Sensing, 2022, 14(20): 5221.

[8] Su S, Shao X, He L, et al. Face image completion method based on parsing features maps. IEEE Journal of Selected Topics in Signal Processing, 2023.

[9] Li J, Li H, Zhu L, et al. Video-Based Sentiment Analysis of International Chinese Education Online Class. International Conference on Computer Science and Education. Singapore: Springer Nature Singapore, 2022: 231-243.

[10] Shen G, Wang X, Duan X, et al. Memor: A dataset for multimodal emotion reasoning in videos. Proceedings of the 28th ACM international conference on multimedia. 2020: 493-502.