Comparison of CNN and ResNet Neural Networks on the Performance of Facial Expression Recognition

Zhiming Gao *

School of Software, Shanghai Jiaotong University, Shanghai, 200240, China

* Corresponding Author Email: 2643917629@sjtu.edu.cn

Abstract. Facial expression recognition is a crucial task in numerous applications, including human-computer interaction, mental health monitoring, and human behavior analysis. Previous studies have primarily focused on individual models or techniques for improving emotion classification accuracy. However, a comparative analysis of different neural network architectures' performance for facial expression recognition is lacking. The main objective of this study is to compare the performance of Convolutional Neural Network (CNN) and Residual Network (ResNet) on the Fer2013 dataset. The author aims to analyze their behavior during the initial training phases and identify the architectural advantages and challenges associated with each model. Both models are trained using the same experimental setup to ensure a fair comparison. The models are trained using the Fer2013 dataset. The author employs a standard protocol for data preprocessing and augmentation. Results show that CNN achieves an accuracy of around 0.5 in the initial stages of training, which is significantly higher than ResNet's accuracy of 0.25. However, as training progresses, ResNet may outperform CNN because it has a more complicated structure that can capture more complex patterns. CNN exhibits superior performance during the initial training stage of the Fer2013 dataset. This reason may lie behind the fact that CNN has a simpler structure which makes it more sensitive on the basic features of the data.

Keywords: Convolutional neural network; deep learning; facial expression recognition.

1. Introduction

Facial expressions objectively express people's inner emotions in interpersonal communication. These years, facial expression recognition (FER) technology has shown important application value in the fields of mental disease diagnosis, fatigue driving monitoring, and security monitoring. In order to enable computers to recognize facial expressions more efficiently and accurately, researchers have been studying related topics.

Before the widespread adoption of neural networks, facial expression recognition was achieved by various feature extraction methods. Typically, the processes used to extract features are often combined with machine learning methods to improve the interpretation and processing of the obtained features. By using methods such as support vector machines (SVM), researchers can use the extracted features to classify different facial expressions [1]. With the prominence of deep learning, neural networks have become the preferred method for classifying tasks. Several deep learning techniques, including MLP [2], are beginning to be used for facial expression recognition tasks to improve accuracy and efficiency. Also in 2016, Bargal, Sarah Adel and others attempted to use another State of The Art (SOTA) model, the 16-layer Visual Geometric Group Network (VGG16), to recognize facial expressions [3].

In order to combine face expression-derived LBP features with focused Features from FAST and rotated BRIEF named ORB, Hernández-Pérez suggested a method [4]. Initially, a facial recognition algorithm was used to every picture in order to extract more valuable features. Secondly, in order to improve computational performance, the ORB and LBP features were extracted from the facial region; specifically, a novel technique called region division was applied in the conventional ORB to avoid feature concentration. Rotation, grayscale, and size modifications had no effect on these properties. The gathered attributes were then classified using a SVM.

A summary of FER research conducted in the last few decades was given in this publication [5]. It beginned with an introduction to conventional facial expression recognition methods before going

on to discuss common FER system types and their primary algorithms. The author then went on to discuss FER techniques based on deep learning, which leveraged deep networks to accomplish "end-to-end" learning.

The application research of multi-level features of convolutional neural networks in FER by Haiduan Ruan et al. provided a data-based model that deliberately combines feature levels to increase the accuracy. The model was tested on the FER2013 dataset and found to have similar performance to the existing state-of-the-art methods [6].

Kaiming He et al. proposed the Residual Network (ResNet) network architecture in 2015[7], which marked a significant turning point in the CNN image, and more than half of the networks used in image classification or its variants have been used for many years until now. More than half of the networks in the field of image classification still used ResNet or its variants. ResNet proposed a residual block structure, which solved the problem to train networks due to gradient disappearance or gradient explosion in deep neural networks, and network performance degradation caused by deeper levels.

2016 saw Kaiming He and his colleagues introduced ResNetV2, a new residual unit that enhances generalization and facilitates training [8]. They present enhanced outcomes with a 200-layer ResNet on ImageNet and a 1001-layer ResNet on CIFAR-10.

In a study that Zhang Qinhu proposed, he first introduced a self-attention mechanism based on the idea of a residual network and computed a weighted average of all location pixels. Subsequently, the concept of channel attention was presented in order to offer distinct choices inside the channel domain and to target the interaction options across several channels. This study's 74.15% accuracy rate on FER2013 demonstrated the value and superiority of this model for option extraction [9].

According to a paper published in 2023, the suggested approach was validated using six different pre-trained CNN models on well-known datasets. The proposed new technique identified Xception-ResNetV2 as the best network and demonstrated remarkable accuracy of 97.58% for a Softmax classifier [10].

The above results show that the ResNet50 model still plays a huge role in the field of FER. This paper will further explore the performance of ResNet neural network in expression recognition based on the comparative study of traditional CNN neural network and ResNet50V2 on FER2013 datasets.

2. Method

This project aims to categorize facial photos according to emotion. Fig 1a) on the left shows the basic CNN architecture the author used in this paper, Fig 1b) on the right showcases the ResNet50V2 architecture.

2.1. Dataset

FER2013 is a benchmark dataset for facial expression recognition, which is widely used in the field of artificial intelligence [11]. It was first introduced by Paul Ekman and Wallace V. Friesen in 1978, and has been widely used in academic research and commercial applications since then.

FER2013 contains a total of 35,887 facial images, including 26,432 different people. Each person in the dataset is photographed with at least one expression, and the expressions include anger, happiness, surprise, fear, sadness, disgust, and neutral as Fig 2 shows. Table 1. Displays the amount of photos and class list in FER2013 the dataset is collected from the Internet, with a wide range of variations in terms of light, angle, pose, and so on.

There are two sets of the FER2013 dataset: a set for training and a set for testing. The training set has 18,592 images, including 6,334 images of 3,167 people with seven expressions. The testing set is further divided into a validation set and a test set, with 8,553 and 8,742 images respectively.

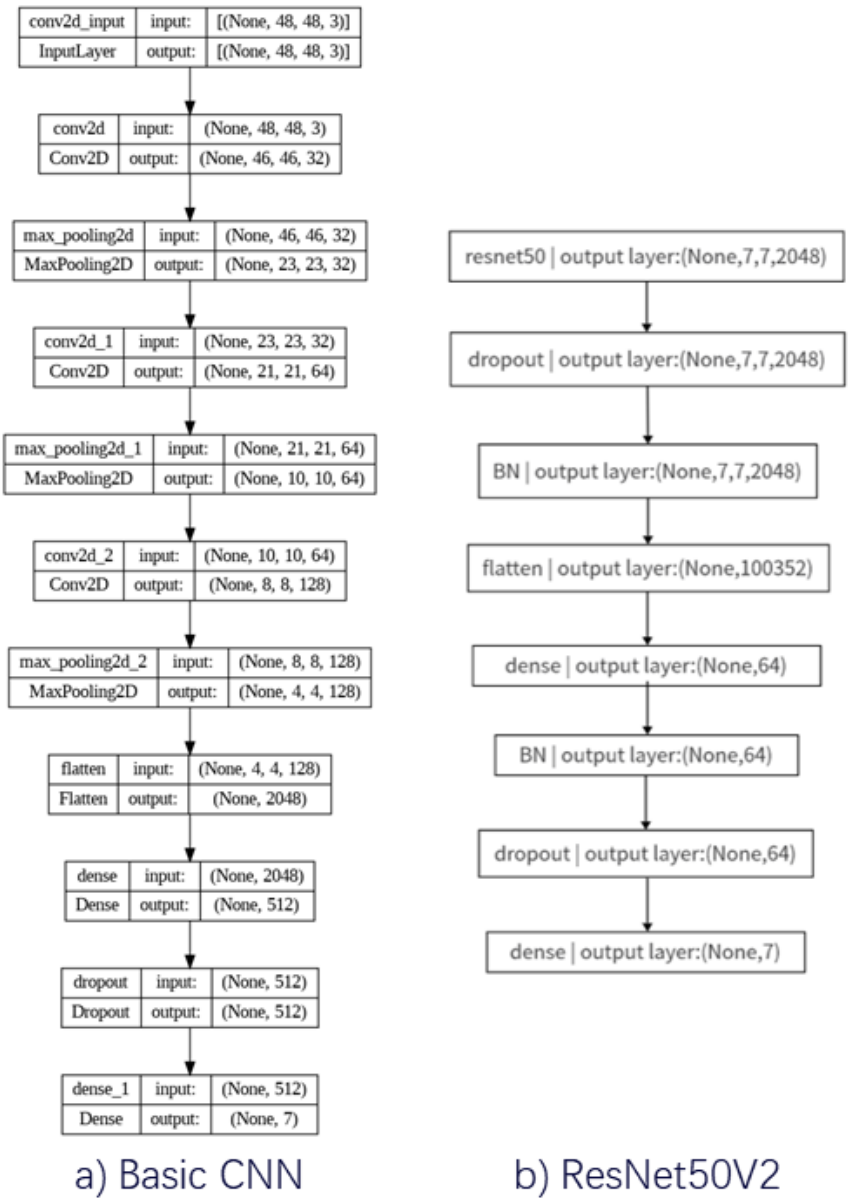

Fig 1. Architecture of basic CNN and ResNet50V2 (Figure Credits: original).

Fig 2. Representative samples from FER2013 dataset [11].

Table 1. Classes and the number of samples in FER2013 dataset.

Class	Number of train images	Number of test images
Surprise	3171	831
Disgust	436	111
Fear	4097	1024
Neutral	4965	1233
Angry	3995	958
Sad	4830	1247
Happy	7215	1774

2.2. Image Preprocessing

For deep neural networks to perform well, a lot of training data is typically needed. Data augmentation can be used to enhance model resilience, prevent overfitting, and expand the quantity of data when there is a restricted supply of data.

The author uses several methods of traditional data argumentation to pre-process images. First, the author rescales the pixel values of the image. I. This step is done to normalize the pixel values and make the input range consistent with the expected range of the model. Second, random rotations are applied to the images.

Third, the author shifts up images to 20% of its width/height in either direction. Fourth, random zoom transformations and shear transformations are used to the images. Finally, horizontal flipping refers to reflecting the image horizontally. This step helps introduce more diversity to the dataset and keep the model from becoming overly dependent on particular image orientations. By applying these transformations to both models, the model is exposed to a larger variety of images, which helps it generalize better and avoid overfitting on the training data.

2.3. Models

2.3.1. Basic CNN model

To develop the CNN architecture, the first step is defining the input layer and the amount of filters in the first convolutional layer. Then, additional convolutional layers are added with increasing numbers of filters, thereafter max-pooling layers to minimize the feature maps' geographical dimensions. After the convolutional layers, fully connected layers are added with ReLU activation to classify the emotions. Different numbers of convolutional layers, filter sizes, and fully connected layers are put into experiments to optimize the model performance. Additionally, this work will use techniques such as dropout and batch normalization to avoid overfitting and improve the generalization capability of the model.

The convolutional layer is one of the basic building blocks of convolutional neural networks in deep learning and is used to extract features from input data. It performs convolution operations on the input data by sliding a learnable convolution kernel (filter) to obtain the feature map. Convolution and full connection are essentially a set of linear transformations. However, compared with full connection, the parameter matrix of convolution is more sparse, and many of the parameters in the kernel matrix are sparse connectivity, while the parameters of the non-zero part are actually parameter sharing.

Pooling operation, similar to the convolution, is also a method of calculating the original image matrix (or feature graph matrix) with a fixed shape window (kernel, or operator) and output the feature map. The maximum pooling layer is used in this study because it can extract the maximum feature data of the specified window, significantly reducing the feature map (the size of the feature tensor), which is the main role of the maximum pooling layer. In addition, since maximum pooling can extract the maximum data for a specific window, regardless of the original location of that data in the window, maximum pooling also alleviates the location sensitivity of the features to be identified.

In deep learning, dropout is a commonly employed technique to address overfitting issues in models. When a model has too many parameters and not enough training data, overfitting of the

trained model is a common occurrence. The model's loss function on the data used for training and elevated prediction accuracy are signs of overfitting. To some extent, regularization effect can be achieved and overfitting can be effectively mitigated by dropout.

Choosing the right loss function is essential to the deep learning models' training process. For this emotion classification task, the classification cross entropy loss function is optimal because it is efficient in multi-class classification problems. Classification cross entropy measures the dissimilarity between predictions labels, and therefore, during the training phase, pushes the model to assign higher probabilities to the correct class. The loss for each image is calculated, where the predicted probability of the model for each emotion class is compared to the actual label of the single thermal coding.

$$H(p,q) = -\sum_i p(i) \log q(i) \quad (1)$$

The above formula is Categorical Cross-entropy loss, where the projected probability of class i is denoted by q(i) and the actual probability of class i is given by p(i). The gradient is then backpropagated through the model to update the weights.

2.3.2. ResNet50V2 model

Although the residual network substitutes the original weighted residual term with identity mapping, it nevertheless borrows the concept of cross-layer connection from the highway network.

The input, x, is transmitted straight through the shortcut connections to the output as the initial result in the residual network structure diagram above. The output result is H(x) =F(x) +x. H(x) =x is the identity mapping that is indicated above when F(x) =0. ResNet modifies the learning objective in this way: instead of learning the entire output, it now learns the "residual," or the difference between the desired value H(X) and x. F(x) = H(x)-x. The next training objective is to estimate the residual result to zero in order to guarantee that accuracy does not decrease as the network becomes deeper.

The output of one layer can immediately cross multiple layers as the input of a subsequent layer thanks to its residual skip structure, which defies the norm that the output of the conventional neural network n-1 layer can only be utilized as the input of the n layer. It is significant because it offers a fresh approach to the issue of the learning model's overall error rate rising rather than falling as several layers are layered.

Neural networks can now comprise hundreds or even more layers, surpassing earlier limitations and enabling sophisticated semantic feature extraction and classification.

ResNet V2 was proposed in the second related publication by the ResNet authors, "Identity Mappings in Deep Residual Networks". Identity Mappings were used in place of the nonlinear activation function (such as ReLU) of the "shortcut connection" in ResNet V2 because the authors discovered that feedforward and feedback signals could be directly transmitted by examining the propagation formula of the residual learning unit in ResNet. ResNet V2 employs Batch Normalization in every layer concurrently. The new residual learning unit has better generalization and is easier to train after this processing.

The structures of ResNetV1 (on the left) and ResNetV2 (on the right) are shown in the Fig 3. The whole ResNet50 is organized as Layer->Block->Stage->Network; Layer is the smallest unit, and ResNet50 represents 50 layers. Block called BottleNeck is composed of three convolution layers. Several blocks are stacked to form a stage. The boxed area in the figure below is a stage. There are 4 stages in a ResNet.

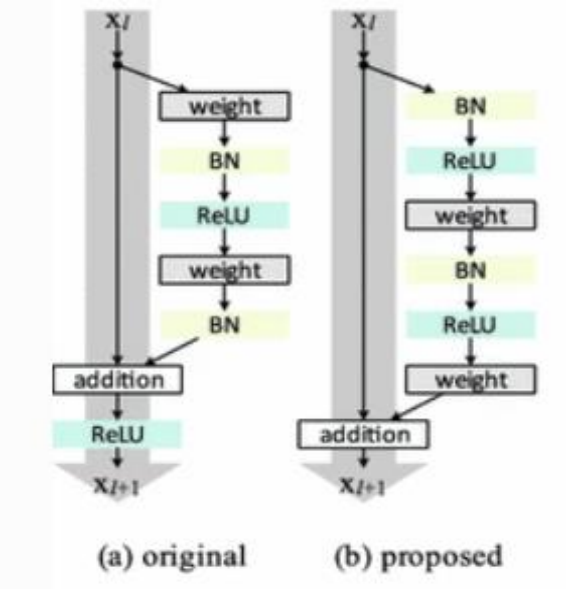

Fig 3. ResNetV1 and ResNetV2 Unit [7].

The input passes through five stages of Resnet50 to the output, with a total of 50 layers. There are two layers (conv7x7, max pooling) in Stage0, nine layers (3x3) in Stage1, 12 layers (3x4) in Stage2, 18 layers (3x6) in Stage3, and nine layers (3x3) in Stage4. Stage0 is relatively simple and can be seen as data preprocessing; the following stages of Stage1, Stage2, Stage3, and Stage4 are all composed of bottlenecks with similar structures.

3. Result

As shown in Fig 4, the basic CNN model achieved a 50% accuracy after only a few hours of training, while the ResNet model exhibited greater volatility in accuracy before stabilizing at a lower final value. In addition, the basic CNN model displayed higher initial accuracy, indicating that it provides a more effective solution for cold start scenarios.

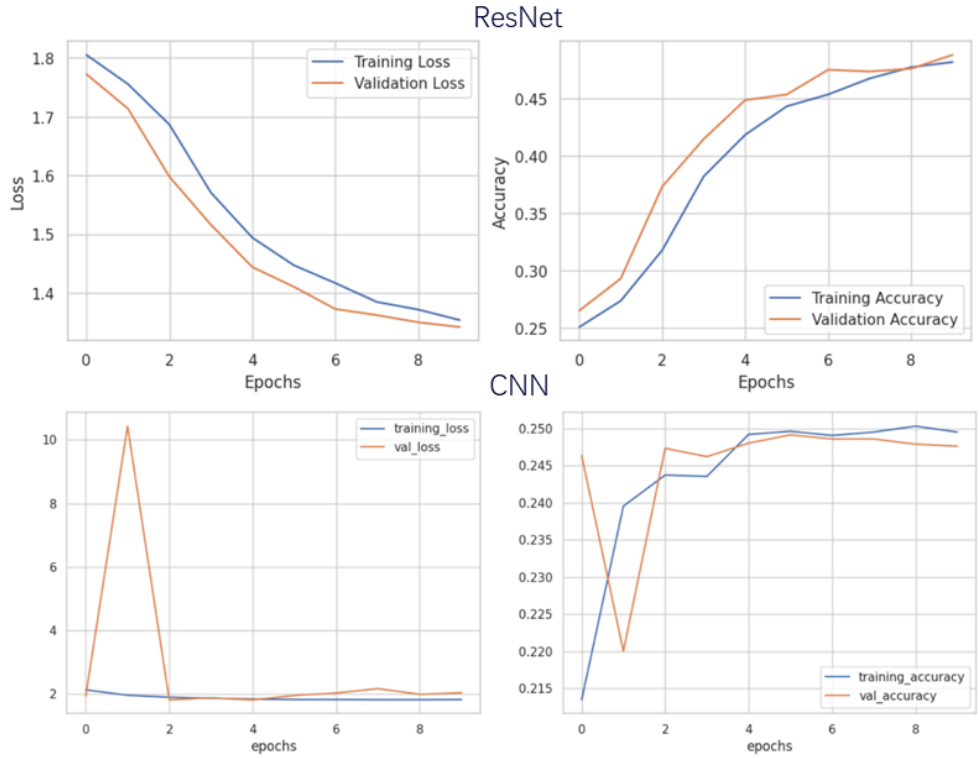

Fig 4. The loss and accuracy of ResNet and CNN models (Figure Credits: Original).

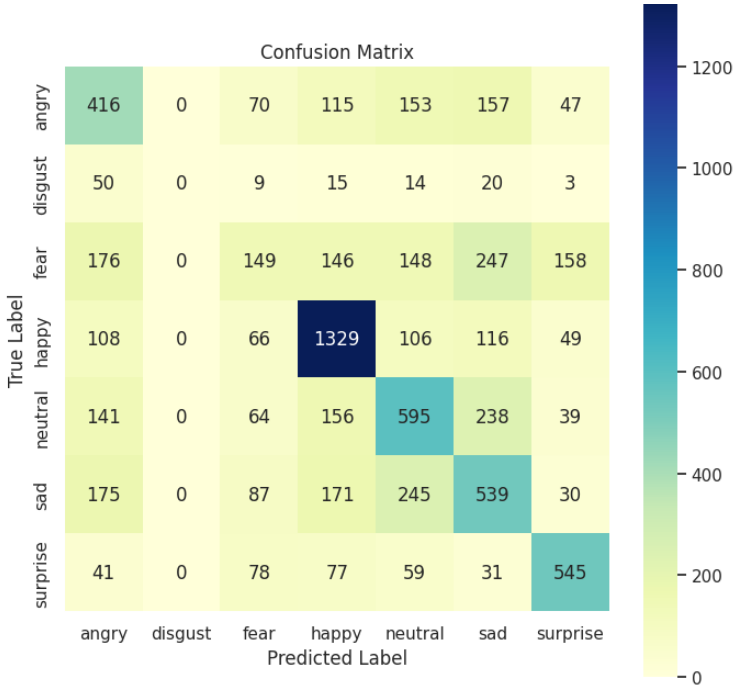

Fig 5. The Confusion Matrix of the basic CNN model (Figure Credits: Original).

The above Fig 5 uses a confusion matrix to provide an explanatory summary of the model performance. The matrix visually represents the distribution of true labels and predicted labels, providing insight into the strengths and weaknesses of the model. Due to the sparse representation in the dataset, the model presents some small challenges in accurately classifying the 'disgust' emotion, as shown and confirmed by the confusion matrix. This observation highlights the need for a more balanced dataset to achieve more consistent emotion classification accuracy across all categories.

4. Discussion

The superior performance of the basic CNN model is attributed to its inverted triangle structure, which avoids the gradient loss too fast in the backpropagation of neural network. Its primary use is the identification of two-dimensional graphs that are distortion-free in terms of scaling, displacement, and other variables. When utilizing CNN, explicit feature extraction is avoided in favor of implicit learning from training data since the feature detection layer gains knowledge from the data. The fact that convolutional networks may learn in parallel due to the fact that all of the neurons on a given feature mapping surface have the same weights gives them a considerable advantage over networks connected by neurons. Due to its particular local weight sharing structure, convolutional neural networks have specific advantages in recognizing spoken words and processing pictures. Its structure is less complex thanks to weight sharing, and it more closely resembles the topology of actual biological brain networks. In particular, the multi-dimensional input vector image can be sent directly into the network, eliminating the need for the time-consuming and arduous reconstruction of data during feature extraction and classification.

However, ResNet50V2 does not exhibit the expected advantage in the initial stage. The accuracy reaches only around 0.25 in the same epoch, which is disappointing. There is also a rare case where the train accuracy is lower than the test accuracy. This may be due to the insufficient number of training epochs. The relatively complex bottleneck structure in ResNet may require more rounds of training for the backpropagation to effectively update the weights. It may also require pre-training to achieve the desired results.

ResNet is a deep network with many parameters, which makes it prone to overfitting the training data. When applying ResNet to the training data, it may learn specific details to the training set rather

than generic emotional expression patterns. A decline of efficiency on the test set may result from this, as the test set may contain different emotional expression patterns compared to the training set.

The FER2013 dataset is relatively small, with only about 3500 images. When using a smaller dataset to train a model, overfitting may occur as the model has more parameters to learn. The introduction of ResNet may exacerbate the overfitting issue, as ResNet has more parameters.

The findings of this study suggest that CNN may be a suitable choice for real-time applications that require accurate emotion classification within a short time frame. On the other hand, ResNet may be preferred for scenarios where a higher level of accuracy is required, such as in critical decision-making processes.

5. Conclusion

In this paper, the author compares the performance of CNN and ResNet neural networks on the Fer2013 dataset. The results shows that CNN exhibited superior performance during the initial stages. This could be attributed to the fact that CNN has a simpler structure, which allows it to focus more on the basic features of the data, while ResNet may take more time to learn and extract complex features. However, as the training progresses, the performance of ResNet may surpass that of CNN, as it has a deeper structure and can capture more complex patterns. This research has implications for the development of facial expression recognition technology. By comparing different neural network architectures, the author provides valuable insights into the design of more effective and efficient models for emotion recognition. Moving forward, advanced training strategies, such as transfer learning or domain adaptation techniques, could be employed to further enhance the performance of these models. Furthermore, it would be interesting to look into how data augmentation methods affect the precision of emotion recognition, particularly in scenarios where labeled data is scarce.

References

[1] Balasubramanian B, Diwan P, Nadar R, Bhatia A. Analysis of facial emotion recognition. In2019 3rd International Conference on Trends in Electronics and Informatics, 2019: 945-949.

[2] Kartali A, Roglić M, Barjaktarović M, Đurić-Jovičić M, Janković MM. Real-time algorithms for facial emotion recognition: A comparison of different approaches. In2018 14th Symposium on Neural Networks and Applications. 2018: 1-4.

[3] Bargal SA, Barsoum E, Ferrer CC, Zhang C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 433-436.

[4] Niu B, GAO Z, Guo B. Facial expression recognition with LBP and ORB features. Computational Intelligence and Neuroscience. 2021, 2021: 1-10.

[5] Ko BC. A brief review of facial emotion recognition based on visual information. Sensors. 2018, 18(2): 401.

[6] Nguyen HD, Yeom S, Lee GS, Yang HJ, Na IS, Kim SH. Facial emotion recognition using an ensemble of multi-level convolutional neural networks. International Journal of Pattern Recognition and Artificial Intelligence. 2019, 33(11): 1940015.

[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 770-778.

[8] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In 14th European Conference of Computer Vision. 2016: 630-645.

[9] Daihong J, Lei D, Jin P. Facial expression recognition based on attention mechanism. Scientific Programming. 2021:1-10.

[10] Sunil MP, SA H. Facial Emotion Recognition using a Modified Deep Convolutional Neural Network Based on the Concatenation of XCEPTION and RESNET50 V2. ResearchGate. 2023: 1-13

[11] FER2013. URL: https://www.kaggle.com/datasets/msambare/fer2013. Last accessed 2023/11/07.