Dean&Francis

ISSN 2959-6157

An Analysis of Facial Expression Recognition Based on the EfficientNet - B4 Model

Hongchun Jiang *

School of Integrated Circuits, Shanghai Jiao Tong University, Shanghai, 201109, China
*Corresponding author: 18341538318@sjtu.edu.cn

Abstract:

Facial expressions are important elements in interpersonal communication. When it comes to human - computer interaction, accurately recognizing users' facial expressions can more effectively understand their emotional tendencies, help to further seize their needs, and give responses that better satisfy their expectations. Making use of deep learning based methods for facial expression identification can play an important role in domains like healthcare and criminal investigation. The study puts forward the idea of using a facial expression recognition model based on EfficientNet - B4. It will be trained with a further improved AffectNet dataset, and the trained model's skill at recognizing test expressions will be evaluated to judge training efficacy. Experimental results show that the current model does an excellent job in classification on the AffectNet dataset. The accuracy of the model improves steadily, and the difference in accuracy between the test and training sets is less than 0. 05.The experiment proves that the EfficientNet - B4 model has a high capacity to deal with the 8 - class emotion classification task of the AffectNet dataset, showing its competence in image recognition and classification. The procedures and conclusions of this study are not only useful as references for related research but also provide valuable ideas for subsequent model refinement and dataset improvement.

Keywords: Facial expressions; EfficientNet - B4; AffectNet dataset.

1. Introduction

The information exchange by means of facial expressions is essential for interpersonal communication. Different expressions can give diverse meanings to the same words [1]. Facial expressions transmit around 55% of the information in face-to-face communication. Interpreting them is very effective in understanding the implied meaning of spoken words. Nevertheless, the variety of facial structures, the in-

tricacy of the brain's neural networks, and possible medical conditions that could impact facial expression display can all present major difficulties to the facial expression recognition process, thus impeding normal communication [2].

With the continuous development of AI, lots of researchers have applied a range of deep - learning models to facial expression recognition. Kong et al. put an innovative MADV - Net algorithm to use for facial micro - expression recognition. On four open - source datasets, it had better test results than 13 mainstream SOTA methods [3]. Yan et al. made use of Emotion - RGC Net for emotion recognition in social media. The results showed that the model could recognize emotions in different social surroundings [4]. Talukder et al. investigated both Support Vector Machine (SVM, a classifier) and dense Convolutional Neural Network (CNN, a feature extractor and classifier) models for facial expression recognition. Subsequently, they proposed a lightweight CNN architecture, which improved the accuracy on their dataset from 63.89% to 78.45%, demonstrating the advantage of streamlined model design [5]. Zhu et al. strengthened the MobileNetV2 neural network, improving recognition accuracy in both test sets and effectively curbing the parameter number [6]. Zhang improved the model by means of a bidirectional attention mechanism and a multi - layer Transformer encoder. The RDA - MTE model he put forward achieved high accuracy across multiple training datasets [7]. Gupta et al. came up with a deep - learning - based way to analyze facial emotions for real - time detection of online learners' engagement, getting 90% accuracy across multiple datasets [8]. Niu et al. put forward an algorithm which combines ORB features (Oriented FAST and Rotated BRIEF) with LBP features (Local Binary Pattern) extracted from facial expressions. An evaluation of the model was done on several challenging databases, and the results showed the framework's effectiveness and accuracy [9].

Traditional CNNs perform single - scale scaling and only change one of the following: width (number of channels), depth (number of network layers), or resolution (input image size). This usually brings about resource waste or accuracy bottlenecks. By contrast, the EfficientNet model introduces the notion of "compound scaling". It simultaneously adjusts width, depth, and resolution in a fixed ratio. This makes it possible to allocate resources more rationally across network layers, so as to fully exploit the performance potential [10]. The study makes use of the open - source AffectNet dataset to train a model relying on the EfficientNet - B4 deep learning network with processed images. The AffectNet dataset, released by Graz University of Technology in 2017, is a large-scale facial expression dataset. It has over 420,000 images. The expression categories cover 8 basic emotions—neutral, happy, sad, surprised, fear, disgust, anger, and contempt—as well as special cases such as no expression, uncertain expression, and no face detected. After the training process, graphs of the accuracy - epoch and loss - epoch functions for the training and test sets are drawn, and images of the confusion matrix are also produced. The study shows that when working on emotion analysis problems, the EfficientNet - B4 model can make certain of emotion recognition accuracy and maintain relatively low computational effort.

2. Research Process and Methodology

2.1 Simple Information of the Dataset

The study makes use of the AffectNet dataset for model training. The dataset consists of 11 categories, e.g., neutral and happy. This experiment is going to use the first 8 image categories for training. With regard to emotional characteristics, the dataset uses emotional valence and arousal to give a more detailed description of emotions. When emotions are made more positive, valence goes higher; when the emotional state is made more excited, arousal goes up. The dataset gives image paths, coordinates of the facial areas within the image collection, and annotation points for cropping facial expressions. The total number of images that the dataset has is 287,580. We split the dataset into training and test sets in a 4 - to - 1 ratio, getting 230,064 training images and 57,516 testing images. The particular data distribution for different emotions is shown in Table 1.

Table 1. Quantity of images related to various feelings

Emotion	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
Training Set	59887	107506	20359	11269	5100	3042	19903	2998
Test Set	14971	26877	5090	2817	1275	760	4976	750

ISSN 2959-6157

2.2 EfficientNet-B4 Model Construction

2.2.1 Compound scaling construction

The EfficientNet model makes use of the same compound coefficient φ to scale depth, width, and resolution all at once. Equations (1) through (3) show the scaling process.

$$depth : d = \alpha^{\varphi} \quad (1)$$

$$width : w = \beta^{\varphi} \quad (2)$$

$$resolution : r = \gamma^{\varphi} \quad (3)$$

Where α, β, and γ are used to represent the depth coefficient, width coefficient, and resolution coefficient respectively. All the parameters are obtained through a minor grid search and satisfy the condition $\alpha * \beta^2 * \gamma^2 \approx 2$, with α, β, and γ all being greater than or equal to 1.

Using the relationship between network width, depth, and resolution, this approach manages to balance scaling across the three dimensions [10].

2.2.2 Depthwise separable convolution

Depthwise separable convolution decomposes traditional convolution into two steps: depthwise convolution and pointwise convolution. When it comes to the "depthwise convolution" process, independent convolutional kernels are given to each input channel, and only intra - channel operations are carried out without cross - channel fusion. The "pointwise convolution" procedure combines all feature maps across channels. By making use of depthwise separable convolution, redundant computations are done away with, FLOPs are lessened, and computational efficiency is improved.

2.2.3 Network architecture

The EfficientNet - B4 architecture includes nine components, six of which are depthwise separable convolution modules. The structure is shown in Fig. 1. As presented in Fig. 1, the input first goes through a 3×3 convolutional kernel (44 filters, stride 2, padding 1) to get initial features. Batch normalization and the Swish activation function are employed to enhance the nonlinearity of these features.

Subsequently, the network goes through six successive convolutional stages, each made up of multiple stacked MBConv6 modules. MBConv6 module expands the number of channels by 6 times through 1 × 1 convolution, extracts features through deep convolution, and finally compresses the channels through 1 × 1 convolution. The number of channels goes up in sequence: 120, 240, 480, 960, 1920, 3840.

After the convolutional steps are done, the network makes use of a 1×1 convolution to reduce the channel number and concentrate on important facts. A pooling layer then condenses the final feature map to a 1792 - dimensional feature vector, remarkably improving translation invariance. Next, this vector is transmitted into a fully connected layer, and the Softmax function gives the class probability distribution. The cross-entropy loss is calculated to adjust parameters and make the network more efficient.

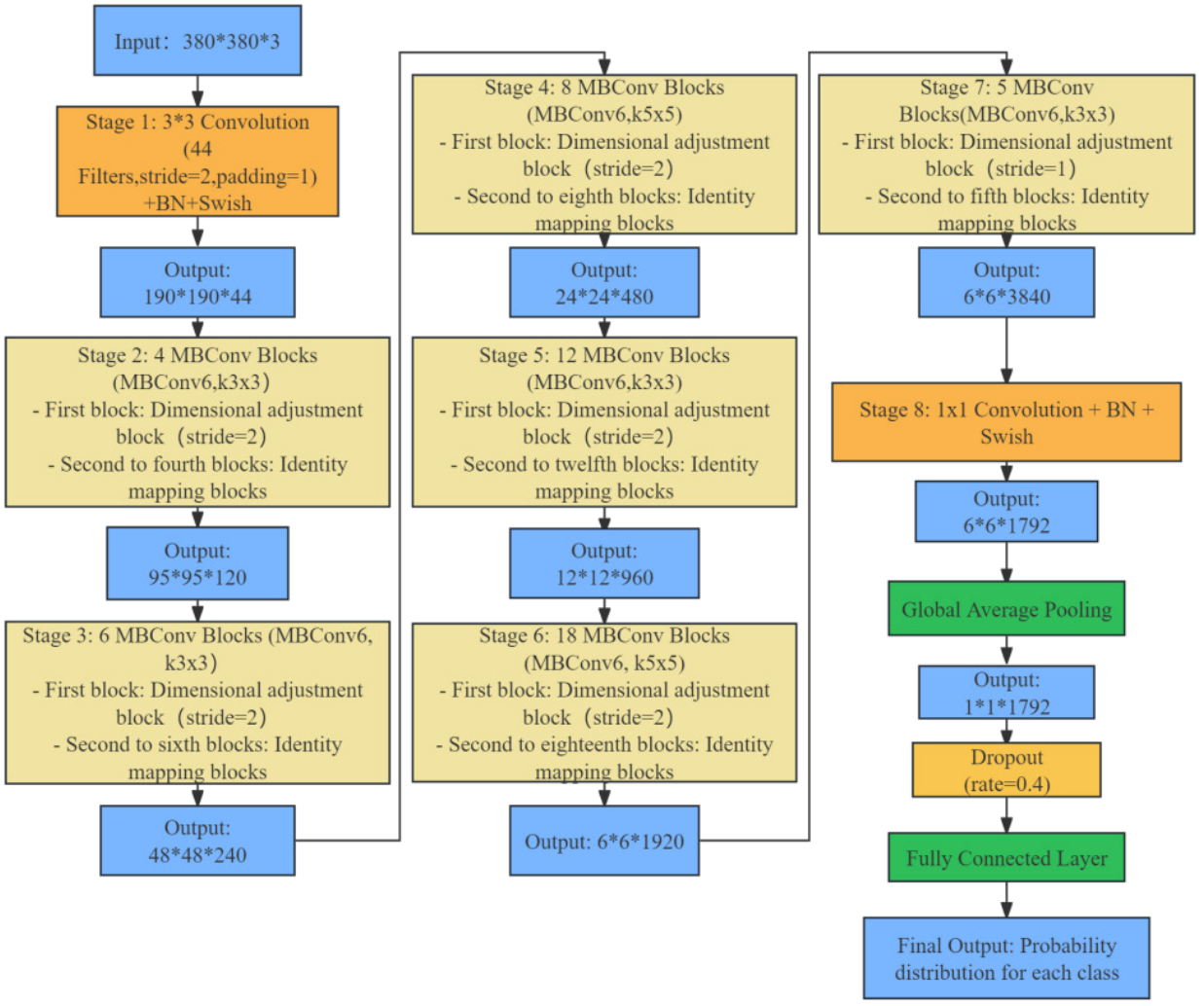


Fig. 1 EfficientNet-B4 Network Architecture

2.3 Training Process

The model was trained with the help of the training dataset for 50 epochs. After each training epoch wrapped up, the model was evaluated by means of the test dataset, and the accuracy and loss values of the test set for each epoch were noted down. After the training is done, the ModelCheckpoint function saves the model parameters with the peak accuracy across all epochs on its own accord. The following shows the calculation process of the test metrics.

To start, the softmax function is used to turn the unnormalized class scores the model gives out into probability scores.

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{4}$$

Where z_i indicates the class score and p_i shows the probability score.

The loss - value expression is shown in Equation (5).

$$loss = -log(p_i) \tag{5}$$

The accuracy can be defined as Equation (6).

$$Accuracy = \frac{Numberofcorrectlypredictedsamples}{Totalnumberofsamples} = \frac{1}{N}\sum_{k=1}^{N}\mathbb{I}(\hat{y}_k = y_k) \tag{6}$$

Were N is the total amount of data included in the assessment, $\hat{y}_k$ is the predicted label of the k - th sample, and y_k is the real - world label.

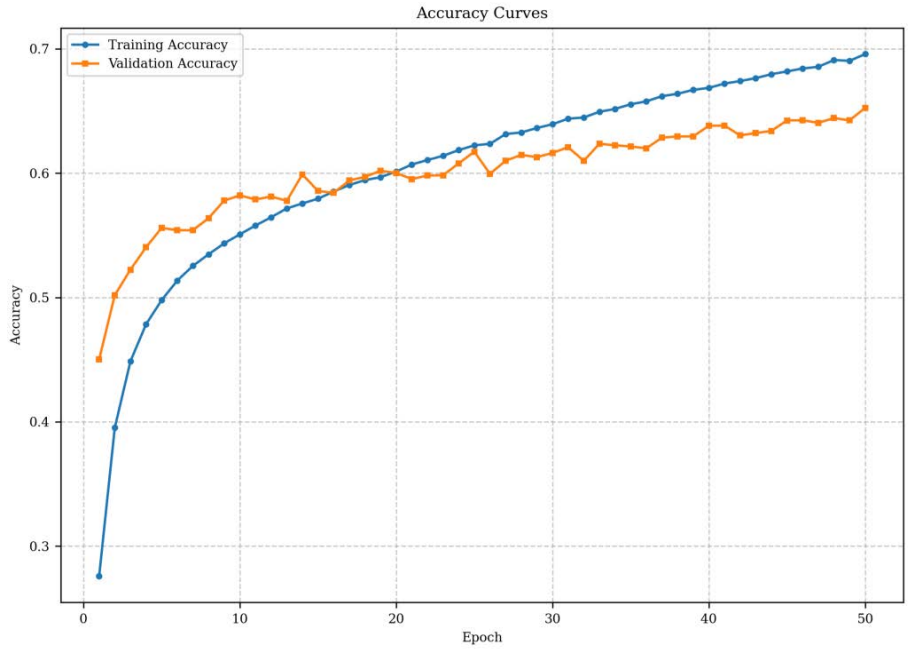
Given that the true labels are known, the confusion matrix shows the types and quantities of samples from one class

ISSN 2959-6157

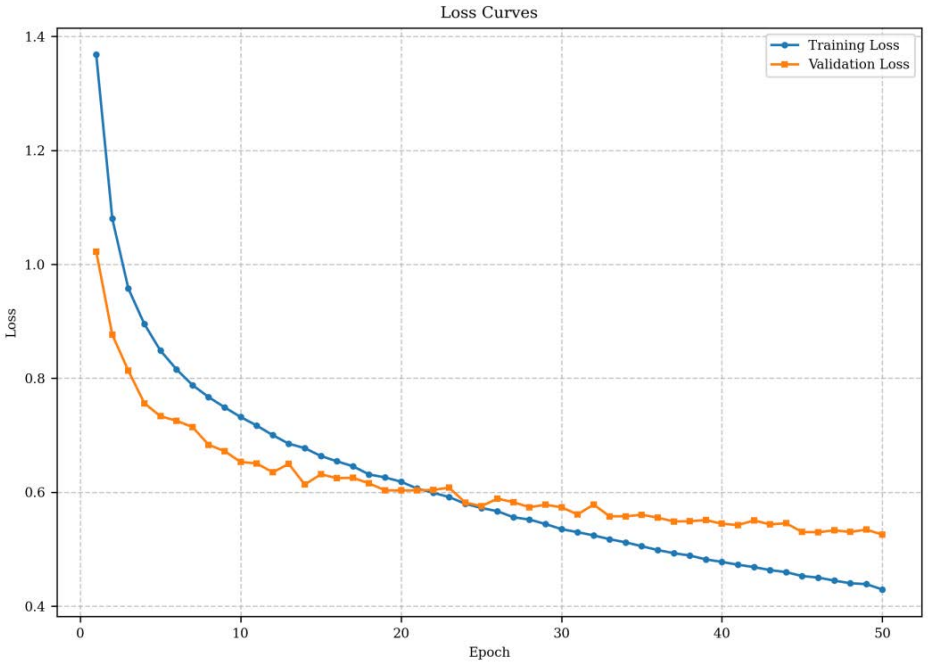
mis - assigned to other classes.

3. Experimental Results

Fig. 2 shows the changes in accuracy and loss values of the training and test sets as training epochs progress. The accuracy of the training set has increased from 0.26 to 0.7, and the accuracy of the test set has risen from 0.45 to 0.65. This indicates that the model learns relevant features from the training set with success. The classification accuracy of training samples gradually goes up, and its fitting ability keeps on strengthening.



(a)



(b)

Fig. 2 Line Charts of Accuracy and Loss: (a) Accuracy Line Chart; (b) Loss Line Chart.

The change in loss values also confirms this conclusion. The training loss shows a stable decrease, starting from an initial value of 1.38 then gradually falling to 0.42 as training epochs go on. It shows that the model can classify the training set better continuously, and the model's classification error on the training set samples continues to decrease, and the prediction accuracy steadily improves. As the training loss does, the test loss starts at 1.02 and shows a decreasing tendency to 0.54 as the training goes on. The pattern of the loss values further verifies the progressive growth of the model's image - classification aptitude. It shows that the trained model has strong abilities in recognizing and classifying facial expressions and good generalization.

4. Summary

The model employed in this paper performs superbly on the training set. With the accuracy constantly high, it shows the model's strong learning ability to fit the training data and extract and analyze image features. The test - set accuracy of the model gradually goes up, the loss value keeps decreasing, and the metrics closely resemble those of the training set. This shows that the classification model can effectively make use of the features picked up from the training set to judge and classify test set data. The experiment results show that training the EfficientNet - B4 model with the AffectNet dataset produces good results, which verifies the model's strong generalization ability for the current task. This study proves the good performance of EfficientNet - B4 when classifying facial expressions. Other analogous studies can build on this experiment to refine the model and dataset and then continue to look into the applications of the EfficientNet - B4 model in the area of facial expression classification. In the same period, using deep - learning networks for facial expression recognition is still of high practicality, as it can play an important part in early detection of related diseases and better human - computer interaction, highlighting its practical value for affect-aware intelligent systems.

References

[1] Niedenthal P. M., Brauer M. Social functionality of human emotion. Annual review of psychology, 2012, 63: 259–285.

[2] Teng C. L., Cong L, Wang W., et al. Disrupted properties of functional brain networks in major depressive disorder during emotional face recognition: an EEG study via graph theory analysis. Front Hum Neurosci. 2024, 18:1338765.

[3] Kong W., You Z., Lv X. 3D Micro-Expression Recognition Based on Adaptive Dynamic Vision. Sensors, 2025, 25(10): 3175.

[4] Yan J., Pu P., Jiang L. Emotion-RGC net: A novel approach for emotion recognition in social media using RoBERTa and Graph Neural Networks. Plos one, 2025, 20(3): e0318524.

[5] Talukder A., Ghosh S. Facial Image expression recognition and prediction system. Scientific reports, 2024, 14(1): 27760.

[6] Zhu Q., Zhuang H., Zhao M., Xu S., Meng, R. A study on expression recognition based on improved mobilenetV2 network. Scientific reports, 2024, 14(1): 8121.

[7] Zhang S. RDA-MTE: an innovative model for emotion recognition in sports behavior decision-making. Frontiers in neuroscience, 2024, 18: 1466013.

[8] Gupta S., Kumar P., Tekchandani R. K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia tools and applications, 2023, 82(8): 11365–11394.

[9] Niu B., Gao Z., Guo B. Facial Expression Recognition with LBP and ORB Features. Computational intelligence and neuroscience, 2021, 2021: 8828245.

[10] Tan M. and Le O. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the36th International Conference on Machine Learning, ICML 2019, Long Beach, 9-15 June 2019, pp. 6105-6114.