Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231152

Facial emotion recognition based on improved ResNet

Wei Du

Faculty of Science and Engineering, University of Nottingham Ningbo China, Zhejiang, 315000, China

ssywd1@nottingham.edu.cn

Abstract. A common and difficult task in computer vision is the identification of facial expressions of emotion. Facial emotion recognition has shown great promise in many areas including healthcare, robotic communication and customer service. However, the variability of human appearance and muscle movement leads to difficulties in facial emotion recognition. Therefore, deeper convolutional neural networks are introduced to recognize facial emotions. The residual network (ResNet) can be built in a deep architecture that can solve the degradation problem when the depth of the network increases. In this paper, an improved ResNet50 is proposed to implement facial emotion recognition. Specifically, the proposed model appends two blocks consisting of fully connected layers to the ResnNet50. The layers are stacked with shortcut connections to solve the degradation problem and to make the training process smoother. The accuracy achieved by the improved model on the Facial Emotion Recognition 2013 dataset (FER-2013) is 13.31% higher than that of the ResNet50. Experimental data indicate that the improved model performs efficiently in facial emotion recognition due to shortcut connectivity and the addition of fully connected layers. Meanwhile, the degradation and gradient disappearance problems are improved.

Keywords: convolutional neural network, residual network, facial emotion recognition.

1. Introduction

Facial emotion is a natural, universal and nonverbal way of expressing an individual's mood. Research has been conducted in various fields such as driver fatigue detection, engagement detection in online education, suicide prevention, and customer emotion analysis. Ekman and Friesen discovered common facial emotions in cross-cultural conditions and proposed the concept of "universal facial emotions" [1]. Basic emotions were defined as anger, happiness, fear, surprise, disgust, and sadness. In 1978, they introduced a standardized coding system tool for analysing movements of facial muscles which consists of 17 action units. Due to its direct and detailed definition of facial emotions, the classification model for discrete emotions dominates over continuous models and the Facial Action Coding System [2].

In the twenty century, traditional shallow learning techniques were used to classify facial expressions. However, in 2004, Feng, Hadid, and Pietikainen introduced a classification scheme based on Local Binary Pattern (LBP) which improved the average correct rate to 77% with the JAFFE database. The LBP operator summarizes the local grayscale relationship between the centre and surrounding pixels as a non-parametric algorithm. LBP-based classification models are fast and accurate because they tolerate variance in monotonic illumination and compute simply [3]. In terms of Non-negative Matrix Factorization decomposition, the components can never be negative. In 2011, Miyakoshi and Kato

DOI: 10.54254/2755-2721/21/20231152

proposed a method that deals with partial occlusion by applying a Bayesian network. A Bayesian network is considered an efficient classifier under uncertain conditions because it modifies the probability distribution with more inspection of samples. The proposed network had a recognition rate of 67.1%, 56.0%, and 49.5% when the image's eyes, brows, and mouths were occluded [4]. Before 2013, datasets were obtained in laboratories where professional actors or researchers expressed facial emotions based on instructions. However, laboratory-controlled images lack complex and real-world scenarios, which causes low accuracy in realistic applications. Since 2013, researchers have found that laboratory-controlled datasets have limitations and lead to large deviations in actual application. To address this issue, more in-the-wild datasets have been developed through emotion classification competitions, such as FER-2013 and the Real-world Affective Faces dataset, enabling real-world applications [5]. In 2020, Zahara et al. proposed a convolutional neural network (CNN) which is inspired by the Xception model, trained by FER-2013 and classifying real-time images captured from a webcam with a mean accuracy of 65.97% [6]. However, the depth of CNNs is limited due to the degradation problem, which causes the accuracy to become saturated and then drop rapidly as the network layers increase. To solve this issue, ResNet was proposed, which eliminates the limitation of depth and increases the accuracy of the model with deeper depths.

The primary objective of the study is to achieve automatic facial emotion recognition. Specifically, first, deep convolutional neural networks are used as the basic backbone for feature extraction. Second, residual blocks are introduced to deal with degradation and gradient disappearance. Third, fully connected layers are used to prompt accuracy by increasing the depth of the architecture. The proposed model has 57.31% accuracy on the FER-2013 dataset, which is 13.31% higher than the classical ResNet50 model with 30 epochs and the Adam optimizer. The experimental results show that the residual structure and the added fully connected layer of the proposed model help to recognize facial emotions efficiently. Meanwhile, the degradation and gradient disappearance problems are improved.

2. Methodology

2.1. Dataset description and preprocessing

FER-2013 is one of the widely used datasets for facial emotion regeneration [7]. FER-2013 consists of 35,887 images separated into 287039 images for training and 3589 for testing and validation. Images of the FER-2013 are grayscale with a resolution of 48x48 pixels. Based on features of facial emotions, the images are classified into six categories: neutral, sad etc. According to the information on images, the labels and images are stored in a Comma-Separated Values (SCV) file. FER-2013 transforms from laboratory-controlled datasets to in-the-wild datasets. Examples in FER-2013 are collected by Google search. Online search ensures diverse image sources and various across-culture scenarios. In addition, images collected online include occluded faces and different resolutions. The data from FER-2013 are considered spontaneous and examples are shown in Figure 1. Unlike ideal or laboratory-controlled examples, the images in the field are closer to the real world. Therefore, models trained by in-the-wild datasets perform higher robustness under a more challenging and stringent reality.

Figure 1. Examples of the FER-2013 dataset.

DOI: 10.54254/2755-2721/21/20231152

2.2. Proposed approach

The model is proposed consisting of the sequential structure of the Resnet50 model, two blocks and a Softmax layer as Figure 2. 48×48×3 inputs go into ResNet50 for feature extraction. Feature maps with a scale size of 2×2×2048 are fed into two blocks for feature mapping. Finally, a Softmax layer is used for multiple classifications. Both blocks consisted of a dropout layer, a batch normalisation layer and an activation layer. Each layer in the first block optimises the performance of the model. dropout layer randomly drops 50 presented neurons to avoid over-fitting. The batch normalisation layer not only reduces the bias of internal covariates but also improves the stability and training speed of the model. The difference is that the first block includes a flattening layer. The flattening layer converts the multidimensional tensor into a one-dimensional vector connecting the fully connected layers.

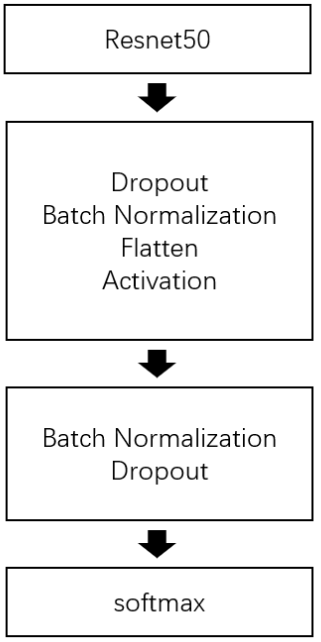

Figure 2. The structure of the improved model.

2.2.1. Resnet. Resnet50 is a type of residual network with a 50-layer architecture. The basic theoretical basis for Resnet is residual learning. It is an explanation of residual learning. $H(x)$ is sat as a fundamental mapping while x is the input of the multiple layers. It is a hypothesis that the multiple layers can approximate the function $F(x) = H(x) - x$. Therefore, the expression of $H(x)$ becomes $H(x) = F(x) + x$. gradually close to zero with the depth of the model increasing and then $H(x)$ equals to x which means the multiple layers achieve identity mapping [8]. By identity mapping, the Resnet is more sensitive to small changes in value. So, Resnet can deal with the degradation problem. In items of the architecture of Resnet, shortcut connections are used for identity mapping. Figure 3 demonstrates the residual structure of Resnet. The 2-layer structure is defined as:

$$H(x) = F(x, \{W_i\}) + x, \tag{1}$$

where x is the input, $H(x)$ is the output of the structure. $F(x, \{W_i\})$ represents the residual mapping for learning and $W_i$is the weight at the layer. As for Figure 3, $F(x, \{W_i\}) = W_2\sigma(W_1x)$ where represents the activation function. σ is the activation function.

DOI: 10.54254/2755-2721/21/20231152

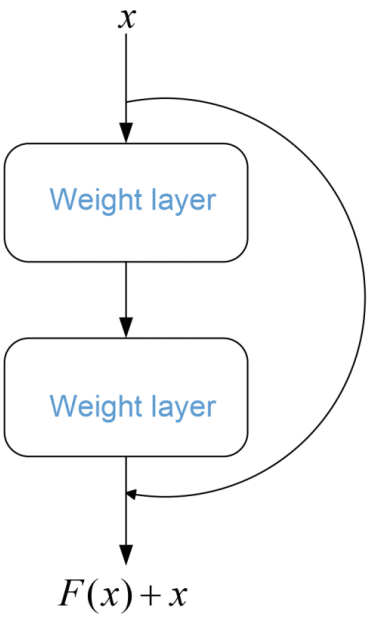

Figure 3. A basic block of Resnet [8].

2.2.2. Optimization. Adaptive Moment Estimation (Adam) is a commonly used approach to descent gradient [9]. Adam optimizer is suitable for objectives with numerous data and non-stationary goals like Root Mean Square Propagation (RMSProp). Adam can operate with squared gradients like Adaptive Gradient (AdaGrad). It is a brief procedure of Adam's work. After initialization, the first and second moment estimates m_t and v_t are updated based on gradients.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{2}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \tag{3}$$

where g_t is the gradient. β_1 and β_2 are exponential decay rates which are default as 0.9 and 0.999. Then, the bias of the first and second moment estimates $\widehat{m_t}$ and $\widehat{v_t}$ are corrected for further computation.

$$\widehat{m_t} = \frac{m_t}{1-\beta_1^t}, \tag{4}$$

$$\widehat{v_t} = \frac{v_t}{1-\beta_2^t}. \tag{5}$$

the Adam update rule:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v_t}}+\epsilon}\widehat{m_t}, \tag{6}$$

where η represents step size, ϵ is a constant equalling 10^{-8}.

2.2.3. Loss evaluation. Cross-entropy loss is popular in the field of CNNs because of the simplicity and efficiency [10]. In the improved ResNet50, the categorical cross-entropy loss is utilized for multiple classifications.

$$J_{cce} = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} y_m^k \times \log\left(h_\theta(x_m, k)\right), \tag{7}$$

where M denotes the quantity of input, K is the quantity of classes. y_m^k represents the true label of learning data m for class k. x is the input and h_θ represents the model with wight.

DOI: 10.54254/2755-2721/21/20231152

3. Result and discussion

Figure 4 shows the trend of the cross-entropy loss and accuracy of the original ResNet50 and the improved ResNet50 with full-connect layers. As for the improved ResNet50, the cross-entropy loss drops from 5.3601 to 1.1556 while the accuracy rises from 41.73% to 57.31%. in terms of the original ResNet50, there are large fluctuations in the loss and accuracy. The loss is larger than 1.4 and the accuracy is less than 45%. Based on the trend of the cross-entropy loss and accuracy, the proposed model has a smoother training procedure. The smooth trend and slight difference between training and validation data in loss and accuracy mean the callback method is suitable for the improved model. So, the performance of the improved model is more efficient than the classical ResNet50.

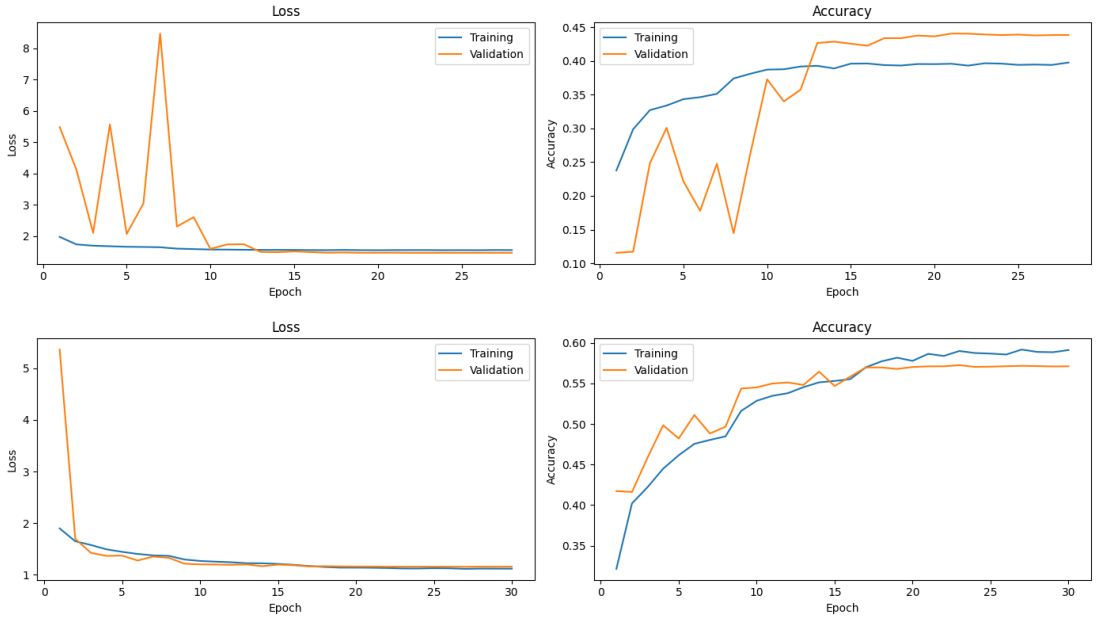

Figure 4. The loss of ResNet50 (upper left), the accuracy of ResNet50 (upper right), the loss of the proposed ResNet50 with fully connected layers (bottom left), the accuracy of the proposed ResNet50 with fully connected layers (bottom right).

To evaluate the effectiveness of the improved model, a comparison is made. Table 1 demonstrates the accuracy of the two models. There is a difference of 13.31%in accuracy between the improved model and the classical ResNet50. Despite having a more complex design than the typical Resnet50, the improved model outperforms the original one. It is because the added blocks consisting of multiple layers construct a deep structure of the model to promote the effectiveness of the improved model.

Table 1. The experimental data of the improved ResNet50 and original ResNet50

Model	Cross-entropy loss	Accuracy
ResNet50	1.47082	44.00%
Improved ResNet50	1.15563	57.31%

Figure 5 is the confusion matrix for the modified ResNet50. The ‘Disgust’ class has a significantly less quantity of data overall than the other six classes. Therefore, the model lacks adequate data to identify the ‘disgust’ emotion. By comparing to the ‘happy’ emotion, the model performs better in recognizing the ‘happy’ emotion because of the sufficient data for the ‘happy’ emotion.

DOI: 10.54254/2755-2721/21/20231152

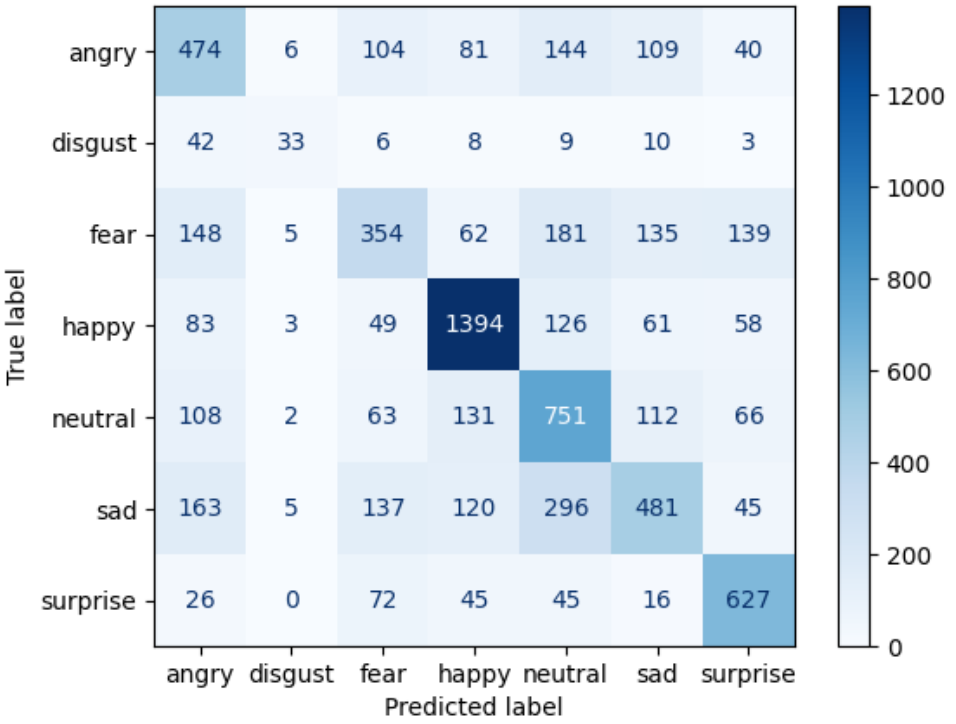

Figure 5. Confusion matrix of the proposed model.

4. Conclusion

This study aims to build a deep CNN for classifying facial emotions. An improved Resnet50 network is used as the backbone for feature extraction. Fully connected layers are stacked with shortcut connections to solve the degradation problem and to smooth the training procedure. The loss and recognition rate of the improved model are compared to those of the traditional ResNet50. Extensive experiments are conducted in the FER-2013 dataset. The accuracy is increased by 13.31% with 30 epochs and the same callback methods. The measured results indicate that the improved architecture is efficient to recognize facial emotion. For further research, deeper Resnet will be trained for more efficient performance. Meanwhile, the residual structure of the ResNet will be redesigned and advanced data augmentation will be considered.

References

[1] Ekman P Friesen W 1971 Constants across cultures in the face and emotion Journal of personality and social psychology 17(2): pp 124–129

[2] Ekman P Friesen W 1978 Facial action coding system Environmental Psychology & Nonverbal Behavior

[3] Feng X Hadid A Pietikäinen M 2004 A coarse-to-fine classification scheme for facial expression recognition Lecture Notes in Computer Science pp 668-675

[4] Miyakoshi Y Kato S 2011 Facial emotion detection considering partial occlusion of face using Bayesian network 2011 IEEE Symposium on Computers & Informatics IEEE pp 96-10

[5] Kusuma G P Jonathan J Lim A 2020 Emotion recognition on fer-2013 face images using fine-tuned vgg-16 Advances in Science, Technology and Engineering Systems Journal 5(6): pp 315-322

[6] Zahara L Musa P Wibowo E Karim I Musa S 2020 The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi 2020 Fifth international conference on informatics and computing (ICIC) IEEE pp 1-9

DOI: 10.54254/2755-2721/21/20231152

[7] Ezerceli Ö Eskil M T 2022 Convolutional Neural Network (CNN) Algorithm Based Facial Emotion Recognition (FER) System for FER-2013 Dataset 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) IEEE pp 1-6
[8] Li B He Y 2018 An improved ResNet based on the adjustable shortcut connections IEEE Access pp 18967-18974
[9] Ruder S 2016 An overview of gradient descent optimization algorithms arXiv preprint arXiv:1609.04747
[10] Ho Y Wookey S 2019 The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling IEEE access pp 4806-4813