A Facial Expression Recognition Model Based on ResNet50

Qingxiao Hua *

King's College London, Department of Mathematics, Strand, London, WC2R 2LS, United Kingdom

Abstract. With the advancement of technology, computers have acquired the ability to recognize human facial expressions from images. This technology can play a crucial role in fields such as psychotherapy, autonomous driving, and public safety. The purpose of this paper is to enhance the accuracy of the computer's classification of facial expressions by applying a neural network to speed up expression-determinant with a high accuracy. Firstly, the project creates a base model uses ResNet50 to pre-train the data on the input 48*48-pixel images. Then, new convolutional blocks with max pooling layer are added to enhance the feature extraction capability. Finally, a global average pooling layer and a fully connected layer with dropout were appended to solve the problem of overfitting. This model performs a classification with 7 different kinds of emotions. This paper employed a Convolution Neural Network (CNN) model and Convolution Neural Network model with residual blocks. And analysis these two models by comparing loss and accuracy. Since the accuracy of both models performs not well, this paper improved the CNN model by appending a ResNet50 network. Eventually, the model based on ResNet50 accomplished the highest efficiency and accuracy which performed best among all the models.

1 Introduction

Humans have various ways to convey information, including sounds, written language, and others. However, the importance of expressions was often ignored by many people; in fact, the amount of information carried by expressions can be extremely large [1]. By identifying expressions, the information obtained can help people make more correct decisions efficiently. Nowadays, there are many scenarios for emotion recognition models. For example, in customs inspections, staff will select those who appear more anxious for further scrutiny. Neuro-medical engineering employed facial Expression Recognition [2]. Also, autonomous cars will be able to recognize the driver's expressions to determine whether the driver is currently in a safe driving state.

The evolution of official-expression recognition has been punctuated by a succession of paradigmatic advances. The earliest handcrafted-feature phase can be traced to Bassili, who pioneered optical-flow-based analysis of official musculature to model dynamic expressions and thus established the critical role of facial cues in decoding fundamental human affect [3].

* Corresponding author: qingxiao.hua@kcl.ac.uk

Extending this line, Padgett and Cottrell combined principal-component analysis with an artificial neural network to implement the first end-to-end representation-learning scheme [4].

The advent of deep learning precipitated a fundamental paradigm shift. Matsugu and Cohen introduced the first five-layer convolutional neural network that jointly optimized features and classifier in an end-to-end fashion, but the hardware of the era precluded practical adoption [5]. A decade later, Tang and Kahou respectively won the FER2013 and EmotiW challenges with a pure CNN and a hybrid CNN-DBN-DAE architecture, unequivocally establishing the superiority of deep learning in accuracy, Nevertheless, the exponential growth in parameters and computational requirements introduced a new bottleneck [6].

For decades, the deep-learning models have been applied in various fields. In particular, the development of computer vision has been very rapid. Now, Computers can recognize features of a large number of objects and learn how to classify them into different groups. Meanwhile, a large number of pre-trained models have emerged as research outcomes. For example, VGG16 convolutional blocks proposed by Simonyan & Zisserman fine-tune the entire architecture with a deep-feature enhanced model to fully preserve its discriminative capacity [7]. In 2023, Li et al. presented FER-former, which integrates a transformer backbone with adversarial debiasing to markedly reduce racial and gender sensitivity, yet inference speed drops to 5 fps and the overfitting problem occurred [8].

This paper has introduced the ResNet50 model designed by Kaiming He, which is a preprocessing model that can greatly prevent overfitting and significantly alleviate the problems of gradient vanishing or explosion [9]. Although human expression is complex, the project has finally trained a model to distinguish among anger, disgust, fear, happiness, sadness, surprise, and neutrality, making the model applicable to people of different cultures and races [10]. It is hoped that this will assist relevant professionals in making more accurate judgments in a shorter period of time. As a contrast, this project also trained two CNN models.

2 Approach

2.1 Feature Transfer

The core objective of this stage is to transfer the general visual knowledge learned from large-scale image datasets to the expression recognition task, addressing the issue of insufficient feature representation capabilities under low-resolution input. This paper used FER2013 as database. Employing the ResNet50 block as its base model, the network adapts the original architecture—originally devised for 48*48*3 RGB image classification—to the 48*48 grayscale emotion domain through three targeted modifications:

1) **Input-channel reduction**. The first convolutional layer is reconfigured from red, green and blue channels to a single colorless channel by applying function 1. This step significantly reduces the risk of overfitting and therefore yields superior adaptation to the limited sample expression dataset while preserving the model's edge-detection capabilities and enhancing robustness to illumination variations.

$$W_{Gray} = \frac{1}{3}\sum_{c=1}^{3} W_{RGB}^{(c)} \tag{1}$$

2) **Pooling-strategy refinement**. Selected pooling layers are removed to preserve spatial resolution throughout the hierarchy. This step prevents excessive compression of the 48×48 input and retains the high-level feature space association.

3) **High-level semantic preservation**. During the initial phase of transfer learning, the model employs a freezing mechanism that locks all pretrained-layer weights, permitting gradient updates exclusively in the next appended layers as the function 2 shown [11]:

$$\frac{\partial \mathcal{L}}{\partial W_i} = 0 \quad , \quad \forall i \in \mathcal{F}_{frozen} \tag{2}$$

For each weight Wi within the frozen layers, the gradient of the loss L remains identically zero. In the backpropagation chain rule, the gradient of the frozen layer is not passed, and its weight remains unchanged during the optimizer's iteration. Due to the fact that the frozen layer does not update the gradient, the computational load of the model's backpropagation is significantly reduced.

2.2 Space Compression

Spatial compression is located after the feature refinement module, whose main goal is to retain discriminative information and eliminate spatial redundancy. This step compresses the two-dimensional feature map into a one-dimensional vector to adapt to the input of the fully connected layer, while avoiding the loss of spatial information induced by global pooling and maintaining the correlation of expression areas. For the 1×1×128 feature cube of the refined output, this model reaches an efficient information condensation through an adaptive compression strategy. (figure 1)

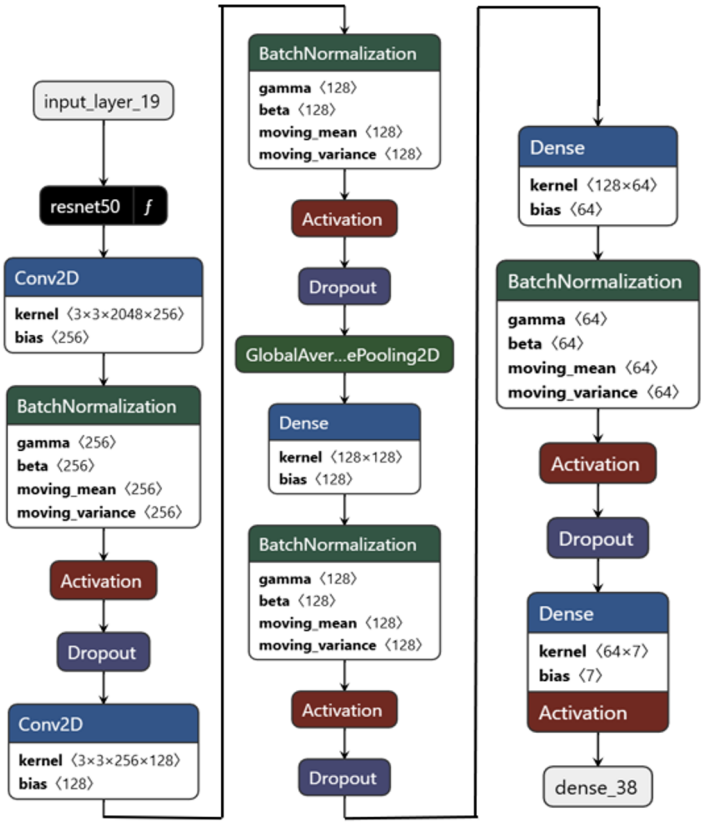


Fig. 1. Model Structure.

2.2.1 Global Average Pooling (GAP)

The global average pooling layer is the core operation of spatial compression, responsible for compressing two-dimensional feature maps into one-dimensional feature vectors [12]. It collapses the feature cube of H × W × C into a 1 × 1 × C vector. The specific mechanism is as follows. By inputting the characteristic cube Fin ∈ R H×W×C, which in this case is H = 1, W = 1, C = 128, the following pooling operations can be carried out.

$$\boldsymbol{v}_c = \frac{1}{\boldsymbol{H}\times\boldsymbol{W}} \sum_{i=1}^{\boldsymbol{H}} \sum_{j=1}^{\boldsymbol{W}} F_{in}(i,j,c) \tag{3}$$

In this function, $\boldsymbol{v}_c$ is the global average response value of channel c and the output vector of the operation is $\mathbf{v} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_C] \in R^C$. In the case of H = 1, W = 1, we have $\boldsymbol{v}_c = \mathrm{F}_{in}(1, 1, c)$, which represents an identity mapping. And in the case of H > 1, W > 1, the spatial information will be compressed into a channel strength distribution. The GAP layer prevents overfitting on the basis of eliminating spatial dimension redundancy and processing next layer with fewer parameters.

2.2.2 Vector Normalization

This process implements standardization and learnable transformation on the feature vectors output by GAP. It first projects the vector onto the unit hypersphere through L2 normalization.

$$\|\boldsymbol{v}\|_2 = \sqrt{\sum_{\boldsymbol{c}=1}^{\boldsymbol{C}} \boldsymbol{v}_{\boldsymbol{c}}^2} \tag{4}$$

$$\boldsymbol{v}_{\boldsymbol{norm}} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \tag{5}$$

Subsequently, scaling and translation are carried out on the spherical surface by placing an affine transformation.

$$\tilde{\mathrm{v}} = \boldsymbol{\gamma} \cdot \boldsymbol{v}_{\boldsymbol{norm}} + \boldsymbol{\beta} \tag{6}$$

γ, β are trainable scaling and offset parameters, which are initialized to 1 and 0 respectively. At this stage, the feature distribution is stabilized from the constraint vector to the unit sphere, while eliminating the differences in feature amplitudes induced by illumination and contrast. At this stage, through a balanced design of spatial compression and information enhancement, the optimization goal of zero spatial dimension and preservation of discriminative information is accomplished, providing high-purity feature vectors for classification decisions.

2.3 Classification

The classification decision-making stage receives the 128-dimensional feature vectors of the spatially compressed output to generate the probability distributions of 7 types of expressions. It contains two fully connected layers and one output layer. The three-layer model design follows the idea of progressive dimensionality reduction, with the dimension decreasing from 128 to 64 and finally outputting in 7 dimensions. The specific methods for the two fully connected layers are as follows:

- global average pooling [12]
- Characteristic nonlinear recombination
- Batch normalization [13]
- ReLU activation
- Dropout [14]

The method employed by our model not only maintains the powerful feature extraction capability of ResNet50 but also significantly enhances the model's classification ability for subtle expressions and cross-domain generalization performance through local feature optimization and regularized intensive design.

3 Experimental Results

This model is trained on 80% of the entire dataset, the rest 20% of the dataset is used as test data. On the training set, we carried out data augmentation processing such as normalization, random rotation, translation, clipping, scaling, flipping, and pixel filling. Normalization processing was carried out respectively on the validation set and the test set. In terms of the configuration of the training environment, we use Adam as the optimizer and select weighted cross-entropy as the loss function:

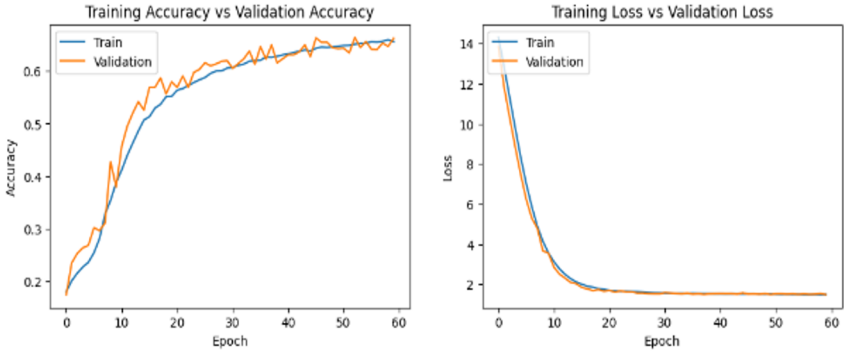
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} W_{y_i} \, log\,(p_{y_i}) \tag{7}$$

For the purpose of alleviating the category imbalance, we set w = [0.1, 0.3, 0.3, 1.0, 0.3, 0.3, 0.1] as the weight vector. The remaining parameter settings are as shown in table 1.

Table 1. Parameter Settings.

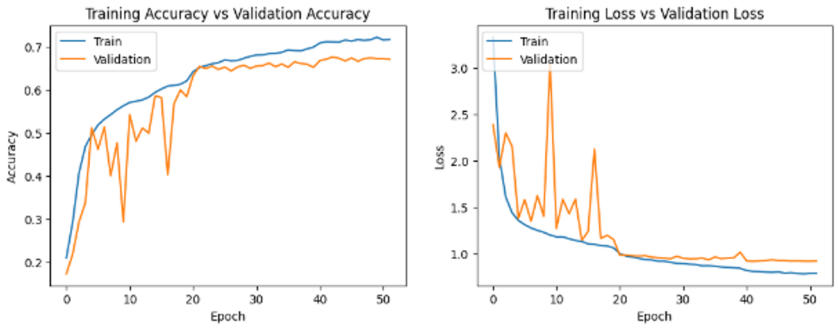
Parameter	Value
Learning Rate	0.0001
β_1	0.9
β_2	0.999
Epoch	60
Batch Size	32

Applied these parameters on the model generalization, the project first gains a Vanilla CNN model with network mainly consists by convolution layers, normalization layers, and max pooling layers. The result of this model has shown in figure 2. The validation accuracy only reached about 68% which is a low value. On the base of Vanilla CNN model, the project builds a second model by remove some convolution blocks to simplify the network and add some residual blocks to skip some blocks, which will reduce the risk of gradient vanishing or gradient explosion. Therefor, this model is better than Vanilla CNN as graph shown in figure 3. However, the validation accuracy of this model is around 70%, still not good enough.



(a) History of Accuracy. (b) History of Loss.

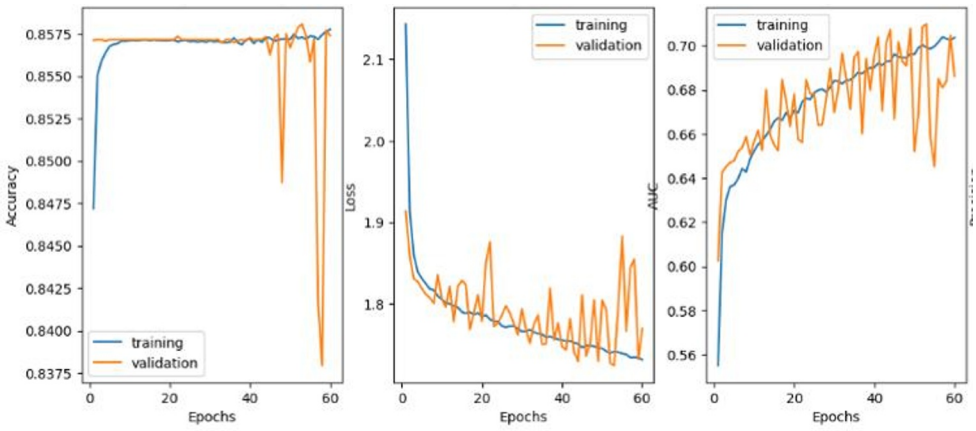
Fig. 2. Accuracy and loss of CNN.



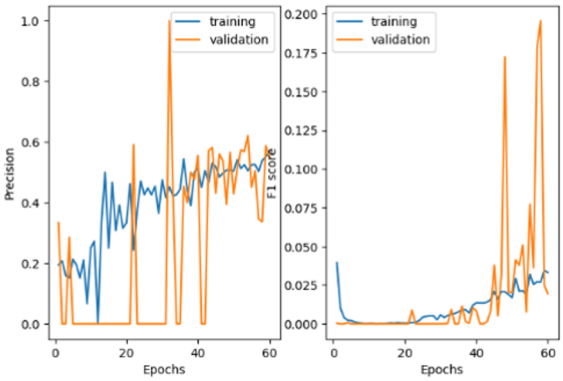
(a) History of Accuracy. (b) History of Loss.

Fig. 3. Accuracy and loss of Vanilla CNN

In the model comparison experiment phase, we first selected three different neural networks and conducted comparison experiments under the two indicators of model accuracy and loss function. The experimental results of model accuracy in the training set and the test set are summarized in table 2. All of these data were collected from figure 2, 3 and 4.



(a)History of Accuracy (b) History of Loss (c) History of AUC



(d)History of Precision (e)History of F1-score

Fig. 4. Results of ResNet50

Table 2. Accuracy Record

Model	Train Acc	Val Acc
Vanilla CNN	0.682	0.662
Residual block + CNN	0.724	0.681
ResNet5 + CNN	0.857	0.857

It is obvious that the accuracy of ResNet50+CNN is much higher than that of the other two models. Arrived at a validation accuracy of 85.7%. These results fully show the advantage of the ResNet50+CNN model working in facial expression recognition. The experimental part used a test set of 48 × 48-pixel grayscale face images to compare the performance of three different neural network models. The comparison results show that ResNet50 model has a significant advantage in the facial expression recognition task, with an accuracy rate of 86% in the test set, which is 16% higher than that of the benchmark CNN. Meanwhile, the Batch Normalization layer increases the convergence speed of the network by 31%, reducing the risk of overfitting while significantly enhancing the stability of the network.

4 Conclusion

4.1 Advantages

This study proposes a model capable of accurately classifying human facial expressions into seven distinct categories. The architecture retains the pretrained ResNet50 backbone and introduces a freezing mechanism during the feature-transfer phase. Subsequently, a two-stage refinement module is appended, in which 1 × 1 convolutions reorganize channel information, while every layer is equipped with a block that contains Batch Normalization, ReLu Activation and Dropout to suppress overfitting.

4.2 Shortage & future work

Although the ResNet50 & CNN model has achieved an ideal overall accuracy rate in the facial expression recognition task, the experimental results show that this model still has limitations. Firstly, the potential of the model has not completely discovered. Keeping

training is helpful while doing future work. Secondly, the current architecture only processes static images and is unable to capture the inter-keyframe features during the evolution of expressions. Changing data base to a radio is a method to solve it. In addition, the balance between the computational efficiency and accuracy of the model still needs to be optimized. In order to figure out this problem, new mathematical optimizers are necessary.

References

1. K. Thomas, S. Vassilios, V. Nicholas, and D. Petros. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. Information, 15(3):135. (2024)
2. D. Pavel, M. Francis, W. Desheng, and G. Claudine. Review of objective topographic facial nerve evaluation methods. Otology & Neurotology, 20(5):672–678. (1999)
3. B. John. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of personality and social psychology. J Pers Soc Psychol 37 (1979)
4. P. Curtis and C. Garrison. Representing face images for emotion classification. Advances in neural information processing systems, 9. (1996)
5. M. Masakazu, M. Katsuhiko, M. Yusuke, and K. Yuji. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural networks, 16 (2003)
6. Y. Jingwei, Z. Wenming, C. Zhen, T. Chuangao, Z. Tong, Z. Yuan, and S. Ning. Multi-clue fusion for emotion recognition in the wild. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, New York, USA (2016)
7. S. Karen and Z. Andrew. Very deep convolutional networks for large-scale image recognition, pages 3601– 3610. (2021)
8. Y. Dingkang, Y. Kun, L. Mingcheng, W. Shunli, W. Shuaibing, and Z. Lihua. Robust emotion recognition in context debiasing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12447–12457. (2024)
9. H. Kaiming, Z. Xiangyu, R. Shaoqing and S. Jian. Deep Residual Learning for Image Recognition. IEEE (2015)
10. E. Paul and C. Daniel. What is meant by calling emotions basic.(Emotion Review, 2011)
11. D. Jeff, J. Yangqing, V. Oriol, H. Judy, Z. Ning, T. Eric and D. Trevo. Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning, pages 647–655. (2014)
12. H. Andrew, S. Mark, C. Grace, C. Liang-Chieh, C. Bo, T. Mingxing, W. Weijun, Z. Yukun, P. Ruoming, V. Vijay. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1314–1324. (2019)
13. I. Sergey and S. Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. (2015)
14. H. Geoffrey, S. Nitish, K. Alex, S. Ilya, and S. Ruslan. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science (2012)