Expression Recognition Based on Multi-level Multi-model Fusion Deep Convolutional Neural Network

Jianzhe Cao [1,†], Chengxin Hu [2,*,†], Lingtong Kong [3,†], Zehao Yu [4,†]

[1] School of International Education, Wuhan University of Technology, Wuhan, China

[2.]Department of Communication Engineering, International College, Chongqing University of Posts and Telecommunications, Chongqing, China

[3.]Department of Mechanics and Engineering Sciences, Shanghai University, Shanghai, China

[4.]Data Science and Big Data Technology, Wenzhou University, Wenzhou, China

* Corresponding Author Email: 1944380@brunel.ac.uk

[†]These authors contributed equally.

Abstract. In the research of Facial Expression Recognition (FER), training networks using deep learning methods often relies on data sets with accurate label and balanced number of each classification. However, when training on low-quality data sets, the performance of traditional models is insufficient. In order to solve the problem, the paper proposes an expression recognition method based on Multi-level Multi-model Fusion Deep Convolutional Neural Network. The Deep Convolutional Neural Network is composed of three models, ResNet-50, ResNet-101 and InceptionNet-V3, which have similar performance on ImageNet. We use the weighted average fusion algorithm as our fusion method. According to the accuracy of the three models, weighted fusion is carried out, and the model with high accuracy is given higher weight. It increases the width of the network and the number of parameters. At the same time, it can solve the problem of accuracy degradation when the number of layers in the deep network increases. We choose FER2013, a widely recognized low-quality data set, for the experiment. The experimental results based on FER2013 data set show that the facial expression recognition accuracy of this method reach 71.78%, which is significantly better than ResNet-50, ResNet-101 and InceptionNet-V3. Compared with other models, it also verifies the effectiveness of the proposed model in identifying low-quality data sets.

Keywords: Image expression recognition, Convolutional neural network, Multi-level, Multiple model.

1. Introduction

Expression are indicators of emotions projected by humans and other animals from their physical appearance, mostly referring to the state formed by facial muscles and the five senses, such as smiling, crying, etc. Facial expressions are part of human language, a physiological and psychological response that is usually used to convey emotions [1]. The American scholar Paul Ekman in 1971 classified universal human emotions into six categories, namely happiness, anger, surprise, fear, disgust and sadness [2]. The complex human expressions can all be blended with these six expressions.

Facial expression recognition technology is a crossover and integration of research fields such as biology, psychology, image classification, machine learning and computer vision by means of face facial image extraction and expression classification. In recent years, with the rapid development of fields such as machine learning and artificial intelligence and the popularity of smart devices, facial expression recognition has become a hot topic in academia and industry.

Facial expression recognition has a wide range of applications in the fields of safe driving, human-computer interaction, medical assistance, robotics and social media analysis. For example, in smart vehicles, the driver's fatigue can be judged based on his facial expression image information, and the fatigued driver can be alerted by issuing a warning, thus ensuring driving safety; in the business field, the emotion of a person can be judged by facial expression recognition, and then the attitude and opinion tendency of the person can be analyzed, which can be applied to consumer analysis and personalized recommendation for shoppers.

Due to the advantages of deep learning methods, deep learning methods are often used in the research of expression recognition in recent years. The applications of deep learning in expression recognition are mostly based on ResNet, GoogleNet and AlexNet network models. The residual network ResNet proposed by Kaiming He et al. uses cross-layer connections to solve the problem of gradient disappearance and gradient explosion in neural networks [3]. Since then, ResNet has been widely used in the field of FER. Mollahosseini et al. combined AlexNet and GoogleNet model to construct a 7-layer CNN for FER, and obtained a good recognition effect [4]. Lopes et al. combined some specific feature extraction methods with convolutional neural network, and first used preprocessing technology to extract some specific facial features [5]. On the basis of classical networks, Weining Wang et al. proposed a new multi-level deep convolutional neural network framework [6], which comprehensively considered the global and local perspectives and designed five levels: original image, salient subject, color, original image local information and color local information. Finally, the decision fusion method is used to fuse the emotion features of the above five levels, and the final result of sentiment classification is output by an SVM classifier [7]. Although the above networks strengthen the ability of feature extraction, they do not consider the problems of imbalanced number of data sets and low accuracy, which will lead to poor classification effect of the model on low accuracy data sets.

On this basis, this paper conducts a study on image sentiment analysis based on convolutional neural networks from the perspectives of building image sentiment analysis models, hierarchical feature extraction of images, mining of important features affecting sentiment, and alleviating the problem of imbalance between various types of samples in existing sentiment image datasets using convolutional neural networks. We finally propose a multi-level multi-model fusion deep convolutional neural network for facial expression recognition. By integrating the three models with small performance gaps, ResNet-101, ResNet-50 and Inception-V3, the weighted arithmetic average method improves the recognition accuracy of the few sample categories in the data set and reduces the impact of misclassified samples, thus improving the overall classification effect.

2. Method

Since various models performs differently with the same dataset, in the model design, we weighted the three models to fuse: ResNet-101, ResNet-50 [1], Inception-V3. A major advantage of fusing is that the classification accuracy of the models can be further improved after fusion by setting different weights to the three models.

2.1. Data augmentation

In this paper, facial expression dataset FER2013 is selected. We use general data augmentation so that the prediction accuracy is more reflected in the model. FER2013 consists of 35887 images, of which we arbitrarily choose 28709 images as the training set, and the remaining images as the validation set. Figure 1 shows the sample images on the collected dataset.

For the training set, the random cropping is firstly performed and the resolution of the cropped image is 299x299. Then a random horizontal flip is performed with p=0.5. Finally, normalization is performed to convert the pixel values to values between [-1, 1].

Figure 1. Samples of dataset

For the validation set, the resolution of images was resized to 320x320. Then center cropping was performed to obtain resolution size of 299x299. Finally, transforming the individual pixel values in range [-1, 1] with tensor form.

2.2. ResNet

In this paper, we select the ResNet-50 and ResNet-101 models as the backbone [3]. In the ResNet structure, there are main branch and shortcut, the main branch is used for feature extraction, and shortcut is used to prevent gradient explosion and gradient decay to make the network go deeper.

For the basic block of ResNet-50 and ResNet-101 [3], a 1x1, 3x3 convolutional kernel is used in the main branch part, which is a bottleneck structure. The first convolutional kernel in the main branch is a 1x1 convolutional kernel, which is used to change the number of channels, i.e., to reduce the number of channels and the amount of computation for subsequent feature extraction. The second convolutional kernel is 3x3 convolutional kernel, mainly used for feature extraction. The third convolutional kernel is also a 1x1 convolutional kernel used to increase the number of channels.

2.3. InceptionNet-V3

Nowadays Inception-V3 [8] is heavily used in visual neural network, which is characterized by factorized convolutions and aggressive regularization compared to other models. These two features can be used to reduce the computational overhead by expanding the convolution operations between the layers of the neural network. Considering the limitation of computational resources in this study, we used the Inception-V3 model which based on ImageNet pre-training as one of the three underlying network models for multi-model deep convolutional neural networks.

In addition, to accommodate the work on expression classification, we changed the number of neurons of the above three models fully connected layer from 1000 to 7. And other weights remain unchanged.

2.4. Combination

We used the weighted average fusion algorithm as a fusion method shown in Figure 2 for multi-model deep convolutional neural networks.

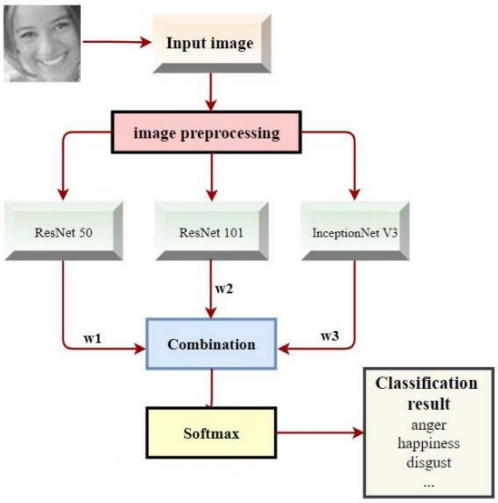

Figure 2. Fusion model

This method can take the three fully connected layer outputs of ResNet-101(y1), ResNet-50(y2) and InceptionNet-V3(y3) to obtains a single output (y0) by weighting and summing the weights (w1, w2, w3) preset for each model shown in formula (1).

$$Output = w_1 * y_1 + w_2 * y_2 + w_3 * y_3 \quad (1)$$

After weighted fusion, the single output y0 will be transformed into a probability distribution of 7 sentiment categories using a softmax classifier. And the weights in the weighted average fusion algorithm will be adjusted by changes in the accuracy of the results. In addition, the training and testing processes of the three underlying networks in the paper were performed independently and the weights were independent of each other.

2.5. Pretrain & fine-tune

Due to the good migration properties of CNN models, a large number of studies have used the method of fine-tuning the pre-trained models on ImageNet to reduce the training time. Therefore, we also provided the corresponding ImageNet-based pre-training weights for the three used base models. And train each of the three base models for 30 epochs against the fully connected layer. Subsequently, in the fine-tuning section, we unlocked the weights of all nodes and trained each of the three base models for 30 epochs. The three base models are trained separately. We wanted to obtain the highest accuracy of each base model in the fine-tuning phase and use it for comparison with the accuracy after fusion.

3. Results and discussion

3.1. Experimental setting

The training of convolutional neural networks is very demanding in terms of graphics card and video memory. In terms of hardware, we used the RTX4000 (16GB) as the training graphics card. And in term of software, Pytorch was chose in this study for implementing the proposed convolutional neural networks. In the actual model training, an epoch can be completed in about 20-30 minutes.

In the model training part, each of the three models uses different batch size and learning rate. In the ResNet-101, the Batch-Size is 8, the learning rate of the pre-training part is 0.001 as well as the learning rate of the fine-tuning part is 0.0001. The Batch-Size and learning rate parameters of ResNet-50 are the same as those of ResNet-101. In the InceptionNet-V3, the Batch-Size of is 16, the learning rate of the pre-training part is 0.001, and the learning rate of the fine-tuning part is 0.00001. All three models use the Adam algorithm for the optimizer and CrossEntropyLoss function for the loss function. Besides, we use accuracy and the corresponding loss as the evaluation criteria.

3.2. Performance of various models

In the face expression classification task, we used the idea of fusion model to build the final network. Specifically, we used ResNet-101, ResNet-50 and Inception-V3 models as the backbone. Finally, the three networks are weighted and fused.

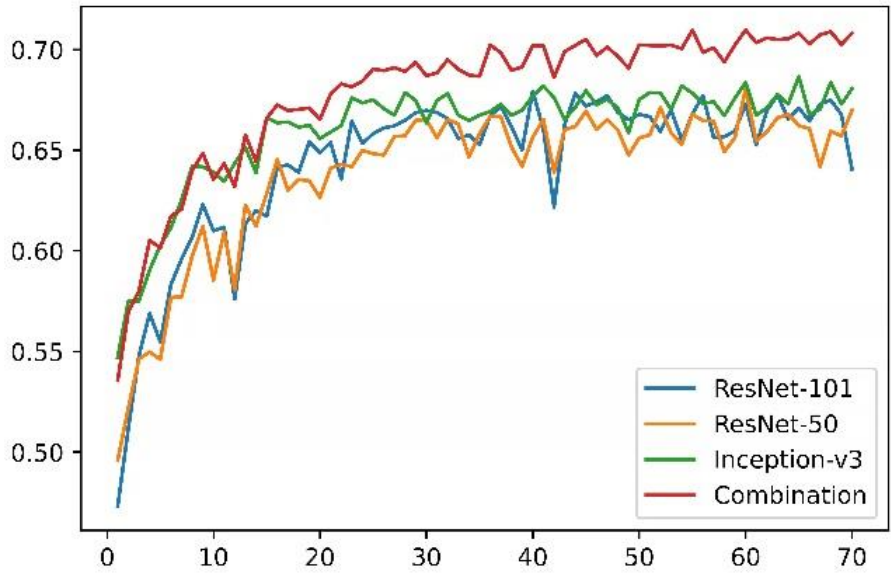

Figure 3. Accuracy curves of ResNet-101, ResNet-50 and Inception-V3

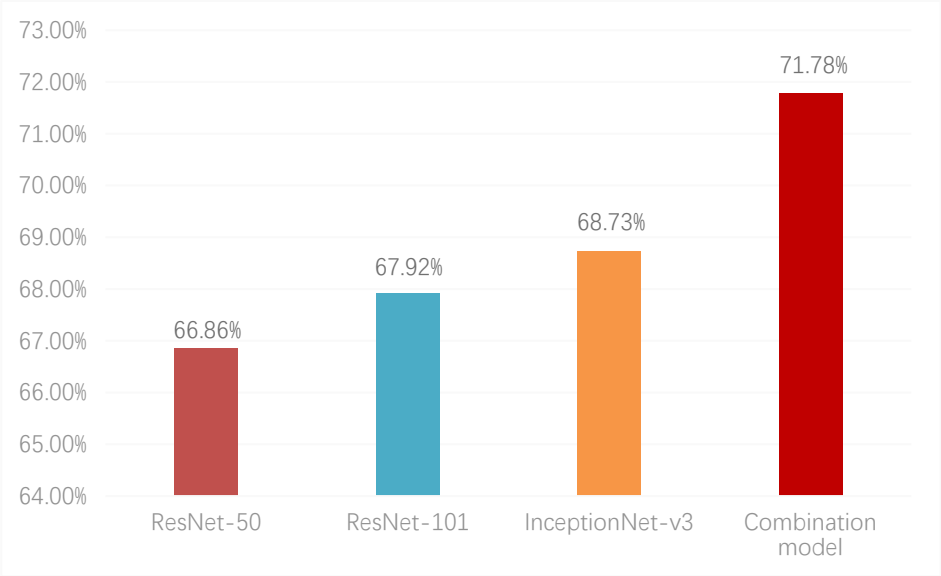

Figure 4. Maximum accuracy of models

Figure 3 and figure 4 demonstrates the effectiveness of the fusion model in this paper. The highest accuracy of the individual model is 68.73% when using ResNet-101, ResNet-50 and InceptionNet-V3 network alone, and after using weighted fusion, the accuracy of the fusion model improves by 3.05% over this accuracy and on average improves by 2.8% over the highest accuracy of individual model.

Table 1. Performance comparison of different methods in FER2013 dataset

Model	Top-1 acc
Combination model	71.78%
ResNet-50	66.86%
ResNet-101	67.92%
InceptionNet-V3	68.73%
VGGNet [9]	60.98%
GoogleNet [9]	63.91%
Improved AlexNet [10]	68.85%

Among the available experimental results shown in Table 1 using FER2013, the first comparison point is VGG and GoogleNet [9], which is 10.8% less accuracy than our combination model. The accuracy of the improved AlexNet is 68.85% [10], which is less than the accuracy of our fusion model.

It is known from the experimental results that the accuracy of the individual models will be lower than that of the fusion model. If the accuracy of the single model is higher, the accuracy of the fusion model will be further improved. In other words, the fusion model obtains the average of the outputs of the three models by weighting the output results and summing them. This result reduces the decrease in accuracy from “bad” results, which in turn will further enhance the increase in accuracy from good results, from where the robustness of the model is improved.

4. Conclusions

In this work, we utilize multiple convolutional neural networks for face expression image recognition. To improve the accuracy of model recognition, we propose a multi-level ensemble deep convolutional neural network framework consisting of three models, ResNet-50, ResNet-101, and InceptionNet-V3. Experimental results show that the accuracy of each sub-model of this framework can outperform existing face expression classification methods in a publicly available face emotion dataset of small order of magnitude. With the fusion model, the accuracy rate will continue to increase

by 2-3%, reaching an ideal recognition rate. In the future, we intend to apply the fusion model to more expression classification tasks and use larger and higher quality image datasets to improve the model accuracy.

References

[1] Ekman P, Friesen W V. Facial action coding system: a technique for the measurement of facial movement [J]. Rivista Di Psichiatria, 1978, 47(2):126-38.

[2] Ekman P et al., Facial Action Coding System, Consulting Psychologists Press,1978

[3] Kaiming H, et al. Deep Residual Learning for Image Recognition [R], arXiv: Computer Vision and Pattern Recognition. 2015

[4] Christian S et al. Going Deeper with Convolutions [R], arXiv: Computer Vision and Pattern Recognition. 2014

[5] Christian S et al. Rethinking the Inception Architecture for Computer Vision [R]. CoRR, 2015, ABS /1512.00567

[6] Wang W et al. Image sentiment classification based on multi-level deep convolutional neural network [J]. Journal of south China university of technology (natural science edition), 2019,47(06):39-50.

[7] Huan R and Pan Y. Decision fusion strategies for SAR image target recognition [J]. IET radar, sonar & navigation, 2011, 5(7): 747-755.

[8] Christian S et al. Rethinking the Inception Architecture for Computer Vision [C], computer vision and pattern recognition. 2016

[9] Gan Y, Facial Expression Recognition Using Convolutional Neural Network[C]// the 2nd International Conference. 2018.

[10] Tan C, Facial expression recognition based on improved alexnet convolutional neural network (in Chinese) [J], Telecommunication technology, 2020, 60 (9): 8