

2203628.Chenyang Li

by Chenyang Li

Submission date: 04-May-2023 02:09AM (UTC+0100)

Submission ID: 204616755

File name: 40779_Chenyang_Li_2203628.Chenyang_Li_498640_230076757.docx (264.36K)

Word count: 2321

Character count: 13186

1. Abstract

The use of computers for facial expression recognition is currently a hot topic in facial expression recognition. Based on deep learning technology, a cascaded classifier is applied to detect and locate the facial region as a whole. A deep convolutional neural network based on self attention mechanism is proposed and used. The model uses Mini-Xception as the basic network and integrates attention mechanism. Then, an expression classification model is constructed by training the convolutional neural network, Finally, achieve fast and accurate facial expression recognition. Several methods are used for experimental comparison in the article, and the final experimental results are analyzed. The results show that the proposed method with the same parameter settings can significantly improve classification performance, recognition accuracy, and real-time speed of facial expression change detection.

2. Introduction

Human expressions are complex and ever-changing, and the expression of joy and anger in facial expressions is one of the most direct and effective ways for people to express their inner emotions. Therefore, research on facial expression recognition technology is unanimously recognized as an important development object for future human-computer emotional interaction [1]. After continuous research and experiments by psychology experts, it has been found that the emotional expressions on human faces can be roughly classified into seven types: disgust, anger, happiness, fear, sadness, and surprise. In actual complex life situations, in order to distinguish from expressionless expressions, a total of seven basic expressions are generally added: normal, As shown in Figure 1. The technology at the forefront of the trend in this era is undoubtedly artificial intelligence. In this field, the facial expression recognition technology proposed in this article has demonstrated extremely high efficiency in various fields such as medical treatment, home furnishing, and criminal investigation. As a current research hotspot, this technology is key to image preprocessing, facial detection, and expression classification [3], as shown in Figure 2.



Figure 1: 7 Common Facial Expressions

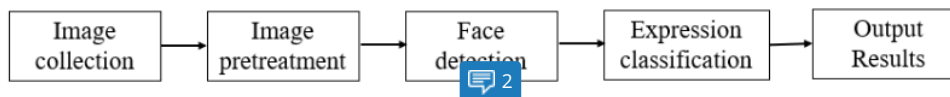


Figure 2: Steps for facial expression recognition

Capture a large number of experimental cases through image collection. In the image preprocessing stage, the size and color of the image are first adjusted to simplify the processing of the input image. Then, through preprocessing, a series of irrelevant factors such as brightness, perspective, and blurry color are eliminated to a certain extent; Then input the results into the input model for training to enhance the accuracy of the

model's recognition. In order to improve the effectiveness of facial expression recognition, the first step is to determine the position of the face in the image. Face detection in the image is to use partial face detection algorithms to detect multiple blocks and scales of the image, and finally obtain the position of one or more people's faces on the object.

In the research process, we first need to process feature engineering, and then use classification algorithms to classify expressions. Although the manual feature extraction method can quickly solve a single simple problem, there are labor costs and fit  3 and other factors to consider for relatively complex situations, and this project is committed to the use of deep learning methods, Constructing an expression classification model by training convolutional neural networks and combining it with attention mechanisms to achieve efficient facial expression recognition.

3. Method

When the system captures an image data, it first uses face detection to determine whether there are faces in the image and the presence of multiple faces. If there are faces, it locates the position of one or more faces. If there are no faces in the image, it returns a prompt message; Then, after capturing facial images on the camera device, Haar features and an Adabsot cascaded classifier are used to classify and detect the captured images. The face detection method in OpenCV and the Haar cascaded classifier used in this article can divide the captured targets into regions for different scales of detection, At the same time, the model proposed in this article based on the Mini Xception network combined with attention mechanism is further used for model training.

3.1. Building a Network Model

Recently, the proposal of attention mechanism has reduced the limitations of computational power and optimization algorithms, making it widely used in various image recognition and classification tasks. This article integrates attention mechanism into the Mini Xception network model and introduces an attention module in the feature extraction section, which has played a certain role in reducing irrelevant interference. By appropriately increasing the number of hidden layers in the network and combining attention mechanism in the feature weighting stage, classification accuracy can be improved, but at the same time, the training time of the model also increases. To effectively solve this problem, the random gradient descent optimizer (SGD) in the gradient descent algorithm (Optimization) is used in this model to reduce the time required to reach the convergence state. Firstly, a model is constructed, which is divided into four parts: convolutional layer, pooling layer, fully connected layer, and output layer. Two convolutional layers and two pooling layers appear alternately in the model, and then 3×3 size convolutional kernels applied to $5 \times$ The input tensor of 5 is used to get the feature map, and the preprocessed expression data is used for model training to finally get the network model. Finally, after activation function processing, a multi-dimensional feature vector can be returned to classify multiple expressions.

3.2. Haar features

The Haar feature is composed of edge features, linear features, center features, and diagonal features, representing the grayscale changes of the image. Figure 3 shows the template used for feature extraction.

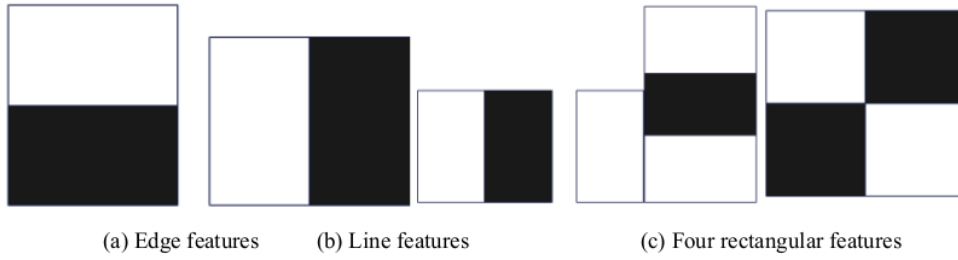


Figure 3 Haar rectangular feature template

The Softmax regression model is relatively similar to Logistic in that it uses the Softmax function for a mapping. The Mini-Xception network architecture [11] further studies expression classification in neural networks by using this layer [12]. The probability of parameter classification in Softmax is

$$p(y^{(i)} = j | x^{(i)}; \theta) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{t=1}^k e^{\theta_t^T x^{(i)}}} \quad (1)$$

In the formula, x is the classification probability, and j is the expression of the classification, θ is the parameter to be fitted.

3.3. Cascading Classifiers

In order to achieve good detection results in OpenCV, several strong classifiers based on AdaBoost algorithm are concatenated to form a concatenated strong classifier. When performing face detection, if all classifiers show the presence of faces in the area, it is determined that there are faces in the area. Otherwise, it is marked as no faces.

In the training model, the applied features will be automatically found by the corresponding algorithm at the threshold, and the face will be divided into two sides. When an error occurs, such as a classification error, the algorithm will automatically select the feature with the lowest relative error rate, which is to separate the difference between the face and the non face. At the beginning of the training, the weight of each image is the same. After multiple classifications, the weight of the incorrect classification will increase again and again, and then continue to execute, Update the weights until the required accuracy or error rate is achieved.

3.3.1. Batch normalization layer

When there are many hierarchical structures in the network, the training convergence speed slows down. The Mini Xception [8] model adds a BN normalization layer between the layer inputs of the network, and the input data from the upper layer is sent to the lower layer after normalization processing. BN layer can control the data distribution of each layer, ensure the stability of data to a certain extent, and further reduce the overfitting problem caused by input changes, which can be expressed as

$$y^{(k)} = \gamma^{(k)} \hat{x}^{(k)} + \beta^{(k)} \quad (2)$$

3.3.2. ReLU activation function

The application of activation function in this paper can effectively strengthen the expression ability of the network model, improve the nonlinear modeling ability of the network [9], and also help the accuracy of neural network classification. ReLu is an unsaturated function, and when its input is positive, the output and

input are linear, without gradient dispersion [10], which can further accelerate the convergence speed.

$$ReLu(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} ReLU(x) = \max(0, x) \quad (3)$$

3.4. Deep separable convolution

If the network model introduced in this article uses common convolution operations, it will greatly increase the number of parameters and training time of the model. Deep separable convolution is to divide the ordinary convolution operation [11] into a point by point convolution and a deep convolution. In the lightweight network model Mini Xception, extracting features will reduce the number of parameters and computational costs.

4. Experimental results and analysis

4.1. Data preprocessing

This experiment used the FER2013 dataset as the dataset. Unlike the experimental dataset, the FER2013 dataset was obtained in a non experimental environment, with a larger dataset and various factors that are more similar to the changes in facial expressions in nature; Taking Sad expressions as an example in this dataset, some samples from the FER2013 dataset are shown in Figure 4. It can be seen that most of the data are not perfectly presented as facial expressions, and are affected by various factors such as light, angle, and limb occlusion. Therefore, this experiment preprocesses the data through image enhancement technology and image normalization.



Figure 4: Example of data samples for Fer2013

4.2. Experimental setup

In order to fully train the model, the experimental settings are as follows:

- (1) The method for data augmentation is as follows: ① normalization processing. ② Effectively perform crop rotation.
- (2) Parameter settings: Horizontal deflection of 0.3, vertical deflection of 0.3, random image rotation of $[-10, 1, 1]$, horizontal flipping of 0, scaling of 0.1.

4.3. Present model experimental results

In the experimental section, the proposed model combined with attention mechanism was compared with other methods on the Fer2013 dataset, and the average accuracy and Epoch time were compared.

It can be seen that the average accuracy and Epoch time of the proposed method reached 95.87% and 1.796 seconds, respectively, which were improved by 1.04% and 0.187 seconds compared to the original basic network. At the same time, the average accuracy and Epoch time were also better than Subnet Ensemble, RestNet-50, and Inception V3 networks, proving the effectiveness of the proposed method. After training, it was found that the accuracy of the seven expressions in the test set was 66%, 67%, 56%, 92%, 60%, and 82%, respectively, 73% of the final calculation shows that the average accuracy of the model proposed in this article on the test set is 70.85%, which has a relatively accurate effect in practical applications.

The average accuracy curve of the training process and the average accuracy curve of the validation set and training set are shown in Figure 5. During each training session, the training set accounts for 75% of the expression image dataset, while the validation set accounts for 25% of the expression image dataset. As the training continues to run, although the training curves fluctuate, the accuracy of the training and validation sets continues to improve until both curves reach a stable state. From Figure 5, it can be seen that the training and testing processes converge quickly, indicating that this method runs faster and has low computational complexity.

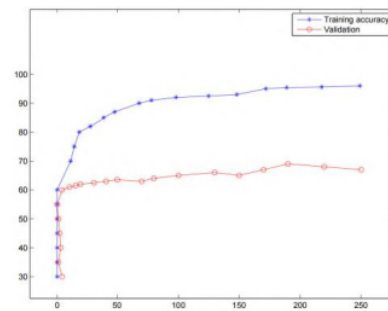


Figure 5: Training and validation curve

5. Conclusion

This article proposes a facial expression image classification model based on Mini-Xception and incorporating attention mechanism for the problem of facial expression classification and recognition. At present, more and more people are deepening their research on facial expression classification. Influenced by attention mechanisms, this article improves the structure of the Mini Xception network model. By adding attention mechanisms and deep separable convolutions, the average accuracy of the model is improved, and the Epoch time is further shortened. Experiments were conducted on the Fer2013 dataset, proving that the proposed network has good classification performance. Compared to the Subnet Ensemble, RestNet-50, Inception V3 methods, the trained network has higher classification accuracy and better real-time performance in this experiment.



Reference

- [1] Qiu Yu, Wang Yanfang Research and Design of a Human Machine Interaction System Based on Expression Recognition [J] Computer CD Software and Applications, 2015 (2): 76-276.
- [2] Liu Zheliang, Zhu Wei, Yuan Ziyang Combining Full Convolutional Networks with Cycle GAN for Image Instance Style Migration [J] Chinese Journal of Image and Graphics, 2019,24(08):1283-1291.
- [3] Fan Xue, Yang Hongbo, Li Yong A Face Image Twisting Algorithm Based on Deep Learning [J] Information and Communication, 2017, 7 (1): 5-9.
- [4] Qian Yongsheng, Shao Jie, Ji Xinxin, et al Multi perspective facial expression recognition based on improved convolutional neural network [J] Computer Engineering and Applications, 2018, 54 (24): 12-19.
- [5] YANG,Y., NEWSAM,S.: Bag-of-visual-words and spatial extensions for land-useclassification[C]. In: Proceedings of Sigspatial International Conference on Advancesin Geographic Information Systems, 2010: 270-279.
- [6] G. CHENG, C. YANG, X. YAO, et al. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs[C]. IEEE Trans. Geosci. Remote Sens., 2018, 56(5): 2811- 2821.
- [7] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 10, pp. 5653–5665, Oct. 2017.
- [8] LAZEBNIK,S., SCHMID, C.,PONCE,J. Beyond bags of features: spatial pyramidmatching for recognizing natural scene categories[C]. In: Proceedings of IEEE Con-ference on Computer Vision and Pattern Recognition, 2006, 2: 2169-2178.
- [9] CHERIYADAT,A.M. Unsupervised feature learning for aerial scene classification[C]. IEEE Trans. Geosci. Remote Sens., 2014, 52(1): 439-451.
- [10] Xu Linlin, Zhang Shumei, Zhao Junli Overview of Image Based Facial Expression Recognition Methods [J] Computer Applications, 2017, 37 (12): 3509-3516, 3546.
- [11] Wang Zhongmin, Li Hena, Zhang Rong, et al Fusion of Convolutional Neural Networks and Support Vector Machines for Expression Recognition [J] Computer Engineering and Design, 2019, 040 (012): 3594-3600.
- [12] BLEI, D.M.,NG, A.Y.,JORDAN,M.I. Latent dirichlet allocation[J]. J. Mach.Learn.Res., 2003, 3: 993-1022.
- [13] LAZEBNIK,S., SCHMID, C.,PONCE,J. Beyond bags of features: spatial pyramidmatching for recognizing natural scene categories[C]. In: Proceedings of IEEE Con-ference on Computer Vision and Pattern Recognition, 2006, 2: 2169-2178.
- [14] CHERIYADAT, A.M. Unsupervised feature learning for aerial scene classification[C]. IEEE Trans. Geosci. Remote Sens., 2014, 52(1): 439-451.

Appendix



CNN_创新.py



CNN_创新2.py

FINAL GRADE

58 / 100

GENERAL COMMENTS

Instructor

Abstract provided but could include some summary results.

You provide an introduction but no real review of related work that could help inform your design and validate your methods. Explain in more detail how related work - eg published models, can inform and be used to guide your choices.

Use a table or similar diagram to easily show your design/architecture.

Results are fair but discuss these further. What about the effect of data imbalance?

Please provide confusion matrix and further analyse misclassifications, etc. Did you consider further dimensionality reduction? Also, please show via bar graphs or further tables the effect of changing parameters in your design. What about overtraining? Why did you train it for so long?

Overall a fair attempt but some room for improvement - see above.



Comment 1

Title should describe what the report is about.



Comment 2

Already assumed present? Expand on this.



Comment 3

could be clearer here.



Comment 4

Adaboost



Comment 5

page break here

GRADING FORM: SMU GENERIC_GJ

CHENYANG LI

0

RESEARCH-INFORMED LITERATURE

KNOWLEDGE AND UNDERSTANDING

ANALYSIS

PRACTICAL APPLICATION AND DEP.

SKILLS FOR PROFESSIONAL PRAC.