A Comparative of Facial Emotion Recognition Performance Based on Cnn, Vggnet And Resnet

Hanting Xue

School of Education, Soochow University, Suzhou, Jiangsu, China

Abstract. With the development of artificial intelligence technology, facial emotion recognition technology has become a research hotspot. The technology has been used in a variety of fields including, but not limited to, human-computer interaction, psychological research and analysis, and security monitoring. In this paper, three classical convolutional neural network models, namely Convolutional Neural Network (CNN), VGGNet and ResNet, are trained using the FER2013 dataset for the task of face emotion recognition, and their generalization ability is verified on FEC Dataset and MMA FACIAL EXPRESSION database. The experimental design covers data preprocessing, model optimization strategies, and standardized training parameter processes. The experimental results indicate that ResNet performs best in cross-dataset tests. Further analysis reveals that the residual connection mechanism effectively mitigates the overfitting problem of the deep network while enhancing the sensitivity to local expression features. The study confirms that model and structure innovation plays a key role in the extraction of complex expression features, provides a model selection basis for emotion recognition in complex scenes, and verifies the superiority of residual structure in cross-domain feature learning, which is an important reference value for dynamic emotion analysis in practical application scenarios.

1 Introduction

Face emotion recognition is an important technology for human-computer interaction and mental health assessment. Face emotion recognition technology can achieve emotional human-computer interaction, allowing machines to better understand human emotions, such as intelligent customer service by recognizing user emotions to provide more targeted services. In the field of mental health assessment, the technology can assist doctors to quickly determine the emotional state of patients and provide a reference basis for diagnosis. The current mainstream methods rely on deep learning, but the performance of different architectures in different scenarios of applications varies significantly.

Khattak et al. proposed an improved convolutional neural network architecture and obtained high accuracy results by optimizing the layer design and parameter configuration [1]. This work not only validated the effectiveness of convolutional neural network (CNN) in multi-task feature learning, but also provides an important reference for emotion

2218401032@stu.suda.edu.cn

computing in complex scenarios, and further promotes the development of human-computer interaction and intelligent perception technology. Minaee et al. proposed an attentional convolutional network (ACN) [2], which combined local feature reinforcement and global regularization strategies, significantly improved the robustness in complex scenes and further promoted the practical development of human-computer interaction and affective computing. Angel et al. proposed a classification framework that combines Inception V3 feature extraction with improved FRCNN, which is a great improvement over traditional methods [3].

In this study, three representative models, CNN [4], VGGNet [5] and ResNet [6], were selected for training using standard dataset FER2013, and two verification sets, FEC and MMA, were used to simulate realistic scenarios, and the performance differences of the models were systematically compared.

In the second chapter of this paper, the basic CNN, VGGNet and ResNet models are briefly introduced. The third chapter first introduces the selected dataset, then introduces the data preprocessing methods, and finally shows the experimental results. Chapter four is about the discussion of this paper. Finally, the fifth chapter summarizes the full text.

2 Method Introduction

2.1 Basic CNN Model

CNN has made remarkable progress in the field of video and image processing, mainly due to its superior feature extraction capabilities. Because of its wide application, CNN has made great progress in these fields. In recent years, more and more researchers have begun to apply CNN to other fields, such as machine translation, sentiment analysis, vehicle detection and speech recognition. In the study of facial emotion recognition, many scholars also use CNN to extract features, and have achieved satisfactory results. The basic CNN is usually composed of the input layer, convolutional layer, activation layer, pooling layer, fully connected layer and output layer. The model used in this paper is a classical CNN architecture of three-stage convolution plus fully connected classification, which is used to identify 7 categories of emotions.

The core function of the input layer is to convert the original image into the normalized data stream that can be processed by the neural network and to assume the dual responsibilities of data preprocessing and format adaptation. The layer reads the RGB image from the path and converts it to the single-channel gray level image, eliminating the color interference and reducing the amount of computation. Enhance image contrast, make dark and bright details more prominent, and alleviate uneven lighting problems, such as the face area blocked by shadows. Retain key information after data is processed.

The first convolutional module is mainly responsible for primary feature extraction and capturing basic visual patterns.64 3*3 convolution kernels are used to scan the input, detect the primary features, and generate 64 different primary feature maps, and the output contains the facial basic structure information, such as the rough position of the eyes and mouth.

The second convolutional module is responsible for intermediate feature extraction, combining primary features to form local semantic units. This layer combines further on the 64 primary feature maps to detect more complex patterns, such as the circular outline of the eye socket and the upward or downward curve of the mouth corner. The output channel is increased to 128, increasing feature diversity. At the same time, continues to introduce nonlinearity, enhance feature separability and recognize local organ morphology, such as narrowed eyes and open mouth.

The third convolutional module is responsible for advanced feature extraction and generates global semantic abstractions. It incorporates intermediate features to capture key combination patterns of expression, such as a combination of wrinkles around the eyes and corners of the mouth for laughter, or a combination of furrowed brows and tight lips for anger. The number of channels is 256, providing a rich combination of features for the fully connected layer. The final size is compressed to 6*6, retaining the most prominent higher-order patterns, and outputs the global semantic information of the encoded expression.

Through the three-level convolution module, the model gradually extracts discriminative high-level expression features from pixel-level data to provide structured information for the classification decision of the full connection.

2.2 VGGNet

VGGNet is a network model proposed by Simonyan and Zisserman. The network structure of VGGNet consists of convolutional blocks composed of convolutional layers and pooling layers stacked. Depending on the configuration of the convolutional kernel and the number of convolutional layers, the VGGNet series contains six different models, which are widely used in the field of computer vision.

The convolutional layer of VGGNet uses a small 3*3 convolutional kernel and uses a fill operation with step size 1. The design of this small convolutional kernel allows the network to learn more local features and to expand the sensory field by stacking them multiple times. Each convolutional layer is followed by a ReLU activation function, which is used to introduce nonlinearity. VGGNet uses a maximum pooling layer of 2*2 to reduce the size of the feature graph while retaining important features. Pooling operations help reduce the computational burden on the network while improving the translation invariance of the network. After the convolution and pooling layers, VGGNet uses several fully connected layers that are used to map the convolutional feature map to the classification output. The final fully connected layer is usually a softmax activation function with a number of classes for multi-class classification.

2.3. ResNet

ResNet is a new kind of network proposed by He et al., that is, a deep residual network. This network won the ImageNetILSVRC champion, and its core structure is the residual module. The core idea of ResNet is to introduce "residual learning", that is, adding a portion of the input to the output of each layer. This allows the network to learn the "residuals" (the difference between the objective function and the input) rather than the objective function itself. In this way, the network can avoid the problem of performance degradation when it is too deep and facilitate the training of the deep network.

With ResNet residuals connected, the network is able to learn the residuals more easily rather than directly learning the objective function, thus improving the performance of the model. Performing well on multiple data sets and tasks can effectively improve the accuracy of tasks such as image classification and object detection.

3 Experimental Results

3.1. Dataset

In this paper, the FER2013 dataset was used for training, and the FEC dataset and MMA dataset were used for verification [7].

Psychologists Ekman and Friesen et al. have proposed six basic human emotions, which are joy, sadness, anger, surprise, disgust, and fear [8]. With the deepening of the research on facial emotion recognition technology, neutral emotion has been formally included in the category of basic emotion, forming the seven basic facial emotions today.

The FeR2013 data set is a facial emotion data set disclosed informing the seven basic Kaggle facial expression recognition competition. The images in the dataset were collected through Google and included partially occluded images of different ages, and different facial angles. Fer2013 is made up of 35,887 grayscale images with a resolution of 48*48, and contains the seven categories of basic emotion labels described above, The dataset consisted of 28,709 images from the training set, 3,589 images from the verification set and 3,589 images from the test set.

Sample examples of various emotional labels are shown in Figure 1.

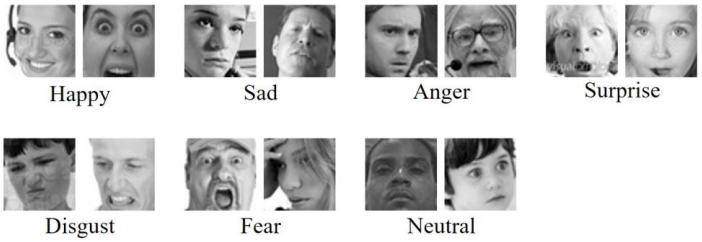

Figure 1 Example of 7 emotion samples in the FER2013 dataset [7]

3.2. Data Preprocessing

Image preprocessing usually begins with gray-scale processing, converting three-channel color images into single-channel gray-scale images, which significantly reduces computational complexity while retaining key texture information.

Then, a normalization operation is performed to linearly scale the original pixel values from 0-255 to the interval [0,1], eliminating dimensional differences to accelerate model convergence [9].

Then the image is uniformly adjusted to the input resolution of 48*48 pixels specified by the model through size normalization to ensure that the tensor dimension matches.

Finally, data enhancement strategies are introduced in the training stage, such as random rotation, translation, horizontal flip and scale transformation, which can expand the sample size, improve the generalization ability of the model and effectively suppress overfitting by generating diverse samples [10].

3.3. Analysis of Results

Table 1 Experimental results

	FER2013		FEC		MMA	
	Accuracy	Time Consuming	Accuracy	Time Consuming	Accuracy	Time Consuming
CNN	99.48%	49.33s	90.61%	64.38s	49.87%	231.80s
VGG	99.77%	14.47s	91.31%	18.28s	54.21%	65.28s
ResNet	99.85%	11.16s	91.65%	14.07s	54.23%	56.21s

As can be seen from Table 1, on the data set FER2013, ResNet performed well, with an accuracy of 99.85% and the shortest time, only 11.16 seconds. In comparison, VGG's accuracy rate is 99.77%, which is only slightly lower than ResNet, and the time is also relatively short, which is 14.47 seconds. CNN, though slightly less accurate than VGG and ResNet at 99.48 percent, had the longest calculation time at 49.33 seconds. As a result, ResNet not only leads in accuracy but also has a clear advantage in computational efficiency.

On the FEC dataset, ResNet again showed its advantage with an accuracy rate of 91.65% and the shortest time of 14.07 seconds.VGG was slightly less accurate at 91.31%, but its calculation time of 18.28 seconds was also much faster than CNN's 64.38 seconds. With an accuracy of 90.61%, CNN was the worst and took the longest to compute, so it performed relatively poorly on this dataset.

Overall, ResNet outperformed the other two models in accuracy and time, making it suitable for this dataset.

On the MMA dataset, the accuracy of VGG and ResNet is close, 54.21% and 54.23% respectively, both of which are better than 49.87% of CNN. Despite similar accuracy rates, ResNet's calculation was more efficient at 56.21 seconds, shorter than VGG's 65.28 seconds, while CNN's was as long as 231.80 seconds. While the overall performance on the MMA dataset was lower, ResNet was still the most balanced in accuracy and computation time.

Overall, ResNet performed well across all three datasets (FER2013, FEC, MMA) with both high accuracy and short computation times, making it the model of choice for a variety of tasks.

4 Discussion

This study draws conclusions by comparing the training effectiveness of CNN, VGG and ResNet on the FER2013 dataset and its cross-scene validation on FEC and MMA databases. With the residual connection design, ResNet50 performs best in the cross-dataset test, which verifies that the deep network can effectively alleviate the gradient disappearance problem through structural innovation, and enhance the ability to capture subtle expressions (such as fear and surprise).

However, the main problem at present is the problem of data bias. Since FER2013 only contains grayscale images there are insufficient samples, which will affect the model recognition rate. At the same time, there are also certain limitations in the coverage of pairs and scenes.Although the generalization ability of the model is verified by using datasets, the datasets do not cover scenes with extreme occlusion (such as masks and sunglasses). Finally, in terms of computational cost limitations, mobile deployment may face hardware adaptation challenges.

At the data level, in the future, people can try to introduce synthetic data techniques such as GAN to generate multi-light, multi-gesture faces to supplement the scarce expression samples Multi-modal data can be integrated to build a multi-dimensional judgment basis for emotion recognition.At the model level,the attention mechanism can be embedded in ResNet to enhance the focus on the key areas of eyes and mouth.And developing lightweight hybrid models that balance the need for speed and precision.

5 Conclusions

With the rapid development of artificial intelligence technology, facial emotion recognition is one of the core tasks in the intersection of computer vision and emotional computing.

With the core goal of improving the ability of cross-dataset generalization, this study systematically compared the training performance of three classical models, CNN, VGGNet and ResNet, on different datasets.

Experimental results show that ResNet, with its residual connection structure and deep network design, achieves the highest accuracy in cross-dataset testing, significantly superior to traditional CNN and VGGNet, and verifies the key role of model architecture innovation in complex expression feature extraction.

However, this study also reveals the current challenges faced by facial emotion recognition technology, such as high computing costs, the need to balance accuracy and efficiency in scenes with strict real-time requirements, and the subjective differences of expression labeling and data distribution offset are still bottlenecks in generalization.This phenomenon indicates that future research needs to improve from the dual dimensions of data and algorithms.

In short, the maturity of facial emotion recognition technology not only depends on the continuous innovation of model structure, but also needs to seek a breakthrough in the triangle balance of data, computing power and ethics.

References

1. A. Khattak, M. Z. Asghar, M. Ali et al., "An efficient deep learning technique for facial emotion recognition," Multimedia Tools Appl., vol. 81, no. 2, pp. 1649–1683, 2022.
2. S. Minaee, M. Minaei and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," Sensors, vol. 21, no. 9, Art. no. 3046, 2021.
3. J. S. Angel, A. DianaAndrushia, T. M. Neebha et al., "Faster Region Convolutional Neural Network (FRCNN) based facial emotion recognition," Comput. Mater. Continua, vol. 79, no. 2, 2024.
4. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint, arXiv:1409.1556, 2014.
6. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
7. I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. Int. Conf. Neural Inf. Process. (ICONIP), Daegu, Korea, 2013, pp. 117–124.
8. P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129, 1971.
9. C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proc. AAAI Conf. Artif. Intell., vol. 31, no. 1, 2017.
10. D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 4489–4497.