Research on Vision Transformer Facial Expression Recognition Algorithm Based on Residual Attention Network

Yuexin Lin*

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, 400074, China
17743355735@163.com
**Corresponding author*

Abstract: *This article proposes a vision Transformer facial expression recognition algorithm based on residual attention network, aiming to improve the accuracy and efficiency of facial expression recognition. By combining the residual network and the attention mechanism, the feature extraction ability is enhanced, enabling the model to capture the subtle features of facial expressions more effectively. The improved Transformer performs extremely well in dealing with complex backgrounds and multi-scale targets, significantly improving the recognition performance. The experimental results show that this method has achieved excellent detection effects on multiple public datasets.*

Keywords: *Residual Attention Network; Vision Transformer; Facial Expression Recognition*

1. Introduction

With the rapid progress of computer technology, people pay more and more attention to the problem of emotional cognition, and the way of emotional expression is constantly exploring and innovating. According to psychologist Albert Mebrabian et al[1], 55% of human emotional expression comes from facial expressions, 38% is aided by language (such as the speed of speech, the frequency and speed of speech, etc.), and the remaining 7% comes from language. Therefore, recognition of facial expressions, as the subject of the external expression of human emotions, is one of the most common, natural and effective ways of many sentiment analysis. At the same time, it has been found that facial expression recognition technology is of great significance in the diagnosis and treatment of depression and autism[2]. However, due to the complex and multi-layered emotions of people in the real world, the computer makes wrong judgments on facial expressions, resulting in the inefficiency of traditional recognition methods in terms of accuracy and reliability.

In recent years, deep learning technology has developed rapidly, especially the convolution-free network structure proposed in the field of natural language processing (NLP), Transformer[3], which has provided new solutions and achieved remarkable performance in the fields of computer vision such as image classification, object detection, and semantic segmentation by drawing on the excellent architectural design in convolutional neural networks (CNNs). Subsequently, the researchers further completed the image classification task on the initial Transformer network application patch sequence, broke the isolation between NLP and CV, and proposed a Vision Transformer[4]. Although ViTransformer has made significant progress in the field of image classification, unlike CNN networks, ViT only considers the global information of the image and does not model the relative relationship between pixels, that is, the local information is lost, so Vit still faces challenges in the application of facial expression image analysis, especially in capturing subtle features. Because facial expression information is affected by factors such as distance, light intensity, occlusion, non-frontal face, multi-target, age, gender, and living environment, and the collected face images are easy to be blurred, the detection accuracy and stability of traditional ViT may decrease when processing such data. To this end, the researchers proposed various improvement strategies, such as combining residual networks and attention mechanisms[5] to enhance feature extraction capabilities, and realizing multi-scale feature fusion through feature pyramid networks (FPNs) [6]. These improvements not only enhance the robustness of Vitransformer in facial expression image analysis, but also provide a new solution for image classification in complex scenes.

In this context, researchers are constantly exploring and improving the performance of methods to enhance recognition. In order to solve the shortcomings of ViTransformer in facial expression recognition, this paper proposes an improved visual Transformer algorithm based on residual attention network. The introduction of residual network can help alleviate the problem of gradient vanishing, effectively transmit deep features through the network, and improve the learning ability of neural networks. At the same time, the application of the attention mechanism enables the model to focus on important feature areas and enhance its ability to capture subtle features.

2. Scheme design and improvement

2.1 Network structure design

The network uses a ResNet residual network and a visual Transformer encoder decoder structure to efficiently learn deep networks. The specific structure is shown in Figure 1:

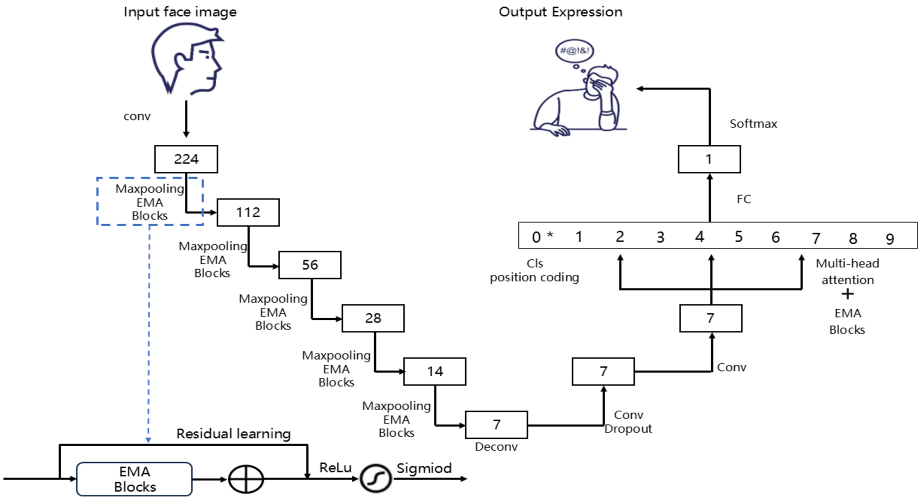

Figure 1: Network structure of facial expression detection

Firstly, the preprocessed image was uniformly cropped to a size of 224×224×3, which contained annotated face expression information, which was input into the ResNet-50 network to extract the local image features of the face, and the feature map was processed through convolution, pooling and residual attention modules, and the deconvoluted features were fused with the low-level features, and then the convolutional layer of Dropout was used to prevent overfitting. Then, the fused feature map is sent to the Vitransformer encoder subnet to extract the global image features, and the feature map is processed by a specific Cls position coding and multi-head residual attention module, and finally, a 1×1×1024 dimensional tensor is output through a fully connected layer and a Softmax loss function, and an anchor frame with a size of 10 is designed according to the distribution of faces. For each anchor frame, there are 5 loss functions, namely the expression category, the confidence level, and the face coordinates.

2.2 Residual network structure design

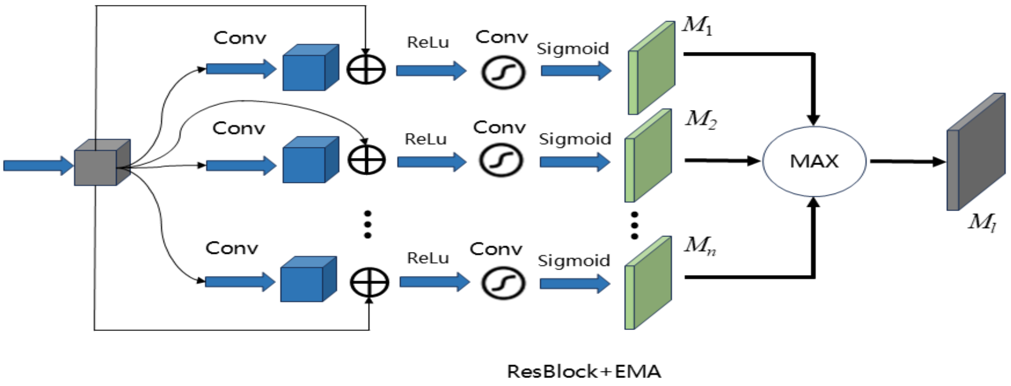

Figure 2: Structure diagram of multi-scale residual attention network

In this article, we designed a feature extraction module that combines a residual network and an attention mechanism, as shown in Figure 2, to improve the accuracy and efficiency of facial expression detection. Residual networks (ResNets) are widely used because of their ability to alleviate the problem of vanishing gradients in deep networks, and the depth and breadth of feature extraction are effectively enhanced through residual connections. However, relying solely on residual networks still has certain limitations when dealing with complex backgrounds and multi-scale targets. Therefore, we introduce a ResNet-based attention module to realize the adaptive adjustment of the weights of the feature map. This improvement allows the network to focus more on the subtle features of facial expressions, which improves the accuracy of detection.

The introduction of the attention mechanism enables the network to dynamically adjust the attention to different features in the process of feature extraction, especially in the detection of local subtle features of faces. By combining channel attention and spatial attention, B local attention matrices are obtained by feeding B local attention networks respectively, and then using element-based maximum operations to aggregate local attention matrices to obtain multi-branch local attention matrix ML. In this article, b is an artificially set parameter, set b= 8. The local attention network consists of two 1×1 convolutional layers, which are used to reduce the dimensionality of the feature map. Networks can effectively filter redundant information and highlight important features. This mechanism is particularly important in human sentiment analysis, because facial expressions are similar to each other, and different types of expressions may only show subtle differences in the same image. Then there is the small similarity between classes, and expressions belonging to the same category may have very different appearances, and these subtle differences are often ignored by traditional methods.

2.3 Improved Vision Transformer

In order to improve the performance of visual Transformer in facial expression detection, a multi-scale residual attention network was introduced in the feature extraction stage. The network combines residual connection and attention mechanism, which can effectively alleviate the problem of gradient vanishing in deep networks, and dynamically adjust the weight of the feature map to improve the attention to important features. This combination method performs well in complex background and multi-scale object detection, and is especially suitable for complex detection tasks such as facial expression classification in real scenes.

Drawing on the successful experience of HIC Vision Transformer in the detection of multi-class cysts in knee joints, the multi-level feature fusion can be further enhanced by adding the multi-scale feature fusion EMA module to achieve multi-level feature fusion, which can further enhance the model's ability to capture the details of facial expressions at different scales. This strategy of multi-scale feature learning and cross-spatial information aggregation can not only capture the local details of expressions (such as smiling at the corners of the eyes or drooping corners of the mouth), but also obtain the linkage relationship between multiple regions (such as the coordination between the eyebrows and the mouth), which has been proved to be effective in improving the detection accuracy of expression recognition tasks, especially in the processing of low-resolution small-sample face images.

Combined with the above improved strategies, the experimental results of the model on multiple public datasets show that its performance in facial expression detection is better than that of traditional methods, which verifies the effectiveness of the residual attention network and the optimized EMA strategy. This study not only provides a new technical means for facial expression detection, but also provides a reference for the improvement of other human sentiment analysis tasks.

2.4 Improved Vision Transformer

In the process of model training and optimization, this study adopts a transfer learning strategy and uses a pre-trained ResNet model as the initialization basis. This strategy not only accelerates the convergence process of the model, but also significantly improves the generalization ability of the model. The application of transfer learning is to make full use of the feature extraction ability of ResNet pre-trained on large-scale datasets, especially in the analysis of negative face-like expressions, which can effectively solve the problem of insufficient data volume. In order to further improve the training effect of the model, data augmentation technology is widely used. Specifically, increasing the diversity of the data through operations such as cropping, rotating, scaling, and flipping can enhance the model's ability to recognize different types of facial expressions. These techniques can effectively simulate different shooting angles and changes in expression morphology when processing face images, thereby improving the robustness of the model. Adjusting the learning rate and using Adam optimizer are critical steps in the optimization process. Dynamic adjustment of the learning rate helps to converge quickly in the early stages of training and fine-tune model parameters in the later stages. The ADAM optimizer further improves the detection accuracy of the model in complex backgrounds through the adaptive learning rate mechanism. Combined with the characteristics of the residual attention mechanism and the improvement of the EMA optimization strategy, the accuracy and efficiency of low-resolution and small-sample image detection are improved. Through the comprehensive application of transfer learning, data augmentation and optimization strategies, this study achieves significant performance improvement in facial expression detection tasks. The combination of these methods not only improves the detection accuracy of the model, but also provides a solid technical foundation for the development of intelligent facial expression recognition system in future cities.

3. Experiments and analysis of results

3.1 Datasets and Experimental Setup

In this study, multiple publicly available facial expression datasets were selected in the experiment, including RAF-DB and AffectNet. These datasets provide rich real-world facial expression data, covering different scenes and categories of facial expressions, which are suitable for evaluating the performance of detection algorithms. To meet the input requirements of the model, the dataset was preprocessed before use, including face alignment, data augmentation, and normalization. Face alignment usually requires face detection, angle correction, and size unification, and data augmentation includes random erasure, color grading, mirroring, and panning, etc., and normalization can ensure data consistency. In the experimental setup, the dataset is divided into training set and test set at the ratio of 80% and 20%, and 5-fold cross-validation is performed to ensure the generalization ability and stability of the model. Cross-validation not only improves the robustness of the model, but also effectively avoids the occurrence of overfitting. The experiments were conducted on NVIDIA GPUs and implemented using the PyTorch framework, which took full advantage of the GPU's computing power and accelerated the model training process.

Table 1 shows the distribution of facial expressions for different categories in the RAF-DB dataset. It can be seen that negative expressions (anger, disgust, fear) account for a small proportion, which puts forward higher requirements for the detection algorithm. In order to address this challenge, this study introduces an attention mechanism into the model and combines CNN and Transformer structures to enhance its ability to recognize small samples.

Table 1: Distribution of facial expressions different types.

facial expressions type	Quantity (piece)	Proportion (%)
Anger	900	6
disgusted	900	6
fearful	450	3
happy	5250	35
neutral	3000	20
sadness	2700	18
surprise	1800	12

The data in the table suggest that detecting small sample expressions is a key challenge for facial expression detection. To this end, the multi-scale EMA fusion feature learning module was used in the experiment, and the feature extraction was optimized through the multi-scale dilated convolution and channel spatial attention mechanism, and its adaptability to complex backgrounds was enhanced.

Experimental results show that the improved Transformer performs well in dealing with complex backgrounds and multi-scale targets, especially in detecting small sample expressions, and achieves significant performance improvements. This result verifies the potential of residual attention network in facial expression detection.

3.2 Evaluate the effectiveness of indicators

In order to comprehensively evaluate the performance of the Vision Transformer based on residual attention network in the detection of different types of facial expressions, this study used 7 categories of facial expressions to detect 7 categories of facial expressions, and the accuracy evaluation index reflected the detection ability and stability of the model for each type of expression. To evaluate the efficiency of the model, the average detection time was also calculated.

3.3 Experimental results and comparative analysis

In experiments, the improved Vision Transformer performed well on the RAF-DB and Affect-Net datasets, especially in detecting expressions in small samples. Table 2 shows a comparison of the accuracy of the two datasets. It can be seen that the accuracy of our method on RAF-DB and Affect-Net datasets reaches 85.4% and 62.2%, respectively. Compared to traditional vision Transformers, the accuracy has been significantly improved, by 5.2% and 3.3%, respectively.

Table 2: Performance comparison between improved Vision Transformer and traditional methods.

Datesets	Method	anger (%)	disgust (%)	fear (%)	Happy (%)	neutral (%)	Sad (%)	Surprise (%)	Acc (%)
RAF-DB	Tradition ViT	77.3	58.7	66.7	87.7	70.3	72.1	74.5	80.2
	Improve ViT	81.6	63.7	72.7	92.4	73.2	77.1	78.7	85.4
Affect-Net	Tradition ViT	58.6	52.5	53.1	85.0	56.0	57.0	61.3	58.9
	Improve ViT	64.9	60.2	61.3	88.1	62.6	62.8	63.7	62.2

Compared with other advanced detection algorithms based on Hyper Faster R-CNN and YOLODM Net, our method also performs well in complex real-world scenarios and multi-scale small-sample object detection. Hyper Faster R-CNN achieves high detection accuracy through multi-scale feature fusion and deep convolution, but there are still some missed detection problems when dealing with small sample facial expressions. The proposed method in this paper significantly enhances the feature extraction ability by introducing the residual attention network and EMA module, especially the detection accuracy in complex real scenes.

In the process of feature extraction, the residual attention network dynamically adjusts the feature weights to effectively filter redundant information and highlight important features. This mechanism has been validated in several studies as an effective means to improve the accuracy of detection. Through comparative analysis, the advantages of the proposed method in feature extraction are further verified, especially in dealing with complex real-world scenes and multi-scale small-sample targets, showing higher robustness and adaptability.

Some of the results of the facial expressions detected by the algorithm in this chapter are shown in Figure 3. The first row is the original unlabeled image, the second row is the image of the real label, and the third row is the results detected by the algorithm in this paper, which are the confidence levels measured in this paper. Each column represents a different type of selected facial expression category from left to right: angry, disgusted, fearful, happy, neutral, sad, surprised. The rectangular box represents the result of the algorithm detected in this chapter, and the number represents the confidence level of predicting the facial expression category.

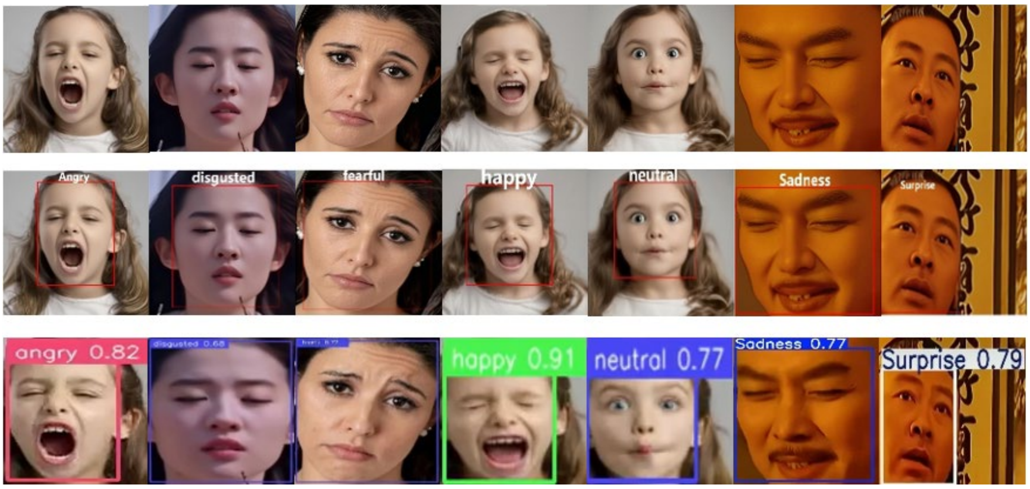

Figure 3: Facial expression detection result chart

Original image, standard diagram of, facial expression, the detection algorithm, in this article

4. Conclusions

The Vision Transformer facial expression detection algorithm based on residual attention network proposed in this article significantly improves the accuracy and efficiency of detection by introducing residual network and EMA attention mechanism, especially in complex real scenes and small sample facial expression detection, which lays a solid foundation for the development of intelligent facial expression recognition system. However, this method still has limitations in dealing with complex and multi-level multi-label facial expressions and occluding non-frontal faces, and the computational complexity is relatively high. Future research can be expanded in the fields of model lightweighting, multimodal data fusion, and cross-domain applications.

References

[1] Mehrabian A. Communication without words [J]. Communication Theory, 2008, 193-200.

[2] Chen Gang, Zhang Shiqing, Zhao Xiaoming. Expression Recognition of Video sequences using Transformer Network [J]. Journal of Image and Graphics, 2022, 27(10):3022-3030.

[3] Li N ,Huang Y ,Wang Z , et al.Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for Facial Expression Recognition[J].Sensors,2024,24(13):4153.

[4] Wang Bin, LIU Tianpei, HUANG Mingliang. Multi-scale high-order reduced bilinear pooling for facial expression recognition[J]. Jiangsu Communications, 2024, 40(05): 98-103.

[5] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[6] Xue Zhichao, Yilihamu Yarmamat, Yan Tianxing. Facial Expression Recognition Based on Multi-scale Feature Fusion of MobileNetV3[J]. Electronic Measurement Technology, 2023, 46(08): 38-44. DOI: 10.19651/j.cnki.emt.2211334.