Fusion Attention Network For Facial Emotion Recognition

Minh-Hai Tran[2], Nhu-Tai Do[1], Tram-Chan Nguyen-Quynh[2], Soo-Hyung Kim [1]
[1] Chonnam National University, Department of Artificial Intelligence Convergence, 77 Yongbong-ro, Gwangju 500-757, Korea
[2] HCMC University of Foreign Language Information Technology, Department of Information Technology,
tranminhhai1506@gmail.com, donhutai@gmail.com, tramtnq@huflit.edu.vn,
shkim@chonnam.ac.kr

Abstract

Deep learning methods and attention mechanisms have been incorporated to improve facial emotion recognition (FER), which has lately attracted a lot of attention. By combining various types of information, the Fusion method has demonstrated an improvement in FER accuracy. In this research, a fusion network with self attention and local attention mechanisms is proposed. It uses a multi-layer perceptron (MLP) network. Using pre-trained models on the RAF-DB dataset, the network extracts distinguishing characteristics from facial images. On the RAD-DB dataset, we outperform the other Fusion methods with quite impressive results.

1. Introduction

Facial exoression recognition (FER) is an important aspect of nonverbal communication, allowing people to express emotions and intentions through facial cues. FER has numerous applications in human-computer interaction, including emotion-based user interfaces, marking research, and psychological studies. However, FER is a challenging problem due to the complexity of facial expressions and the variability in how people express emotions.

As a result, the computer must know and examine specific points on the face and comprehend the traits of each emotion in order to recognize the emotions on the face. Six emotions that are universal across cultures were discovered in one of Paul Ekman's first groundbreaking studies [1].

Additionally, attention mechanisms [2] have been frequently employed for problems including emotion classification in recent years and have produced better results. The Fusion method [3] is also utilized as much as possible in the research, in addition to increasing the model's effectiveness .

In this paper, we propose a technique for combining an MLP network with an attention mechanism to produce significant features from two pre-trained networks.

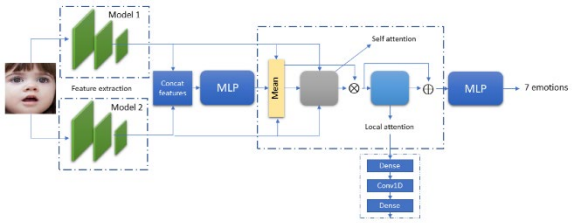
On the RAF-DB [4] dataset, we constructed an attention fusion network and compared it to other fusion techniques. Compared to contemporary methods, we get excellent results.

2. Proposed Method

2.1. Fusion Technical

In this study, we suggest a method that addresses the problem of emotion recognition by fusing two models together. The goal of integrating various models is to maximize each model's benefits and reduce its drawbacks in order to enhance the performance of the recognition of emotions the network.

2.2. Fusion attention Network



(Figure 1) Fusion attention network

Our approach involves creating an attention-based network that includes mechanisms for both Self-attention and Local attention. (as depicted in Figure 1). We employ a fusion method to combine the final features of two emotion recognition models that have already been trained. The goal of this combination is to minimize the weaknesses of each model while maximizing their strengths. To generate a new feature the same size as the input sizes, we first concatenate the two features and pass them through a Multi-layer Perceptron (MLP). We average the features before passing them through the self attention block and the local attention block, and then back and forth through a completely connected network to classify emotions..

2.3. Self Attention and Local Attention

With the goal of self-attention to function, feature values that serve as the input are obtained from two pre-trained networks that act as the Key and Query. Before passing through a softmax function, these features are multiplied by a dot product, then divided by the batch. The feature was

then taken from the MLP and included with the other two input features (shown in Figure 1) through averaging in order to identify the Value features as defined in formula (1).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \otimes V \quad (1)$$

Where K is the key, Q is the Query and V is the Value and is also used in the research [5]

To conduct the convolution operation between the input filter's components, Local Attention employs fully connected layers and a Conv1d layer. For the model to learn and comprehend information from the data more effectively, noise is removed from the data using Conv1d, and abstract feature of the data are created. Additionally, we use feature additions like Resnet to prevent unintended feature loss.

3. Experiments and Results

3.1. Feature extraction networks

We use RAF-DB dataset and pre-training to extract features to perform Fusion method. There are around 30,000 different face images acquired from the internet on RAF-DB. We use the basic set to conduct experiments consisting of 3068 images for testing and 12271 images for training. We train networks like VGG11 [6]. VGG13 [6], Resnet18 [7]. Resnet34 [7] to do feature extraction networks. We train for 50 epochs, batch size 48, learning rate is 0.0001. Also we use augmentation transformations to avoid overfiting such as Fliplr, Rotate in the range [-30,30] and RemoveSaturation.

3.2. Results and Discussion

On the RAF-DB dataset, we train the VGG11, VGG13, ResNet18, and ResNet34 networks for feature extraction. During training, we used Image-Net weight to improve the convergence of the Table 1 gives a description of the concept.

(Table 1) Baseline model for feature extraction

Model	ResNet18	ResNet34	Vgg11	Vgg13
Accuracy	85.3%	85.5%	85.23%	85.2%

(Table 2) Comparison with other fusion methods

Fusion Technical	Model 1	Model 2	Accuracy
Late Fusion	ResNet18	ResNet34	86.35%
	VGG11	VGG13	86.08%
	VGG13	ResNet34	85.98%
	VGG11	ResNet34	86.08%
Early Fusion	Resnet18	ResNet34	86.66%
	VGG13	ResNet34	85.49%
	VGG11	ResNet34	86.08%
Joint Fusion	Resnet18	ResNet34	86.05%
	VGG13	ResNet34	86.63%
	VGG11	ResNet34	86.4%
Our method	ResNet18	ResNet34	90.95%
	VGG13	ResNet34	90.92%

ResNet18 and ResNet34 obtained the highest combined results for Late fusion and Early fusion, respectively, with combined results of 86.35% and 86.66%, as shown in Table 2. VGG13 and ResNet34 have the best results for Joint Fusion 86.63%.

Results from our suggested approach are about 4% better than those from other approaches (shown as table 2). In addition, the model is about 0.4 lower than the initial 90.48% when local attention is not used.

4. Conclusion

The goal of this research is to achieve high efficiency in the Face Emotion Recognition (FER) task using a Fusion network that combines attention mechanisms. A multi-layer perceptron combined with self attention is used by the network to filter out important traits from the two models. Additionally, using the Local attention mechanism helps in verifying features before they pass on to the classifier for classification the seven feelings. On the RAF-DB dataset, the proposed approach's higher accuracy compared to other Fusion approaches has been proven. The process of feature selection is enhanced by the use of local attention, and the model's accuracy is increased. Future research may focus on how effectively the suggested approach works with further datasets and related task.

References

[1] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion.,” *J. Pers. Soc. Psychol.*, vol. 17, no. 2, p. 124, 1971.

[2] M.-H. Guo *et al.*, “Attention Mechanisms in Computer Vision: A Survey,” *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.

[3] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, “Early vs Late Fusion in Multimodal Convolutional Neural Networks,” in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, Jul. 2020, pp. 1–6. doi: 10.23919/FUSION45008.2020.9190246.

[4] S. Li, W. Deng, and J. Du, “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2584–2593. doi: 10.1109/CVPR.2017.277.

[5] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2010.11929

[6] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition.” arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.