얼굴 감정 인식을 위한 로컬 및 글로벌 어텐션 퓨전 네트워크

Minh-Hai Tran[2], Tram-Tran Nguyen Quynh[1,2], Nhu-Tai Do[1], 김수형[1]
[1]전남대학교 인공지능융합학과
[2]베트남 호치민 외국어 정보 기술 대학 정보 기술과
tranminhhai1506@gmail.com, tramtnq@huflit.edu.vn, ntdo@jnu.ac.kr, shkim@jnu.ac.kr

Local and Global Attention Fusion Network For Facial Emotion Recognition

Minh-Hai Tran[2], Tram-Tran Nguyen Quynh[1,2], Nhu-Tai Do[1], Soo-Hyung Kim[1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University
[2]Dept. of Information Technology, HCMC University of Foreign Language Information Technology, Vietnam

Abstract

Deep learning methods and attention mechanisms have been incorporated to improve facial emotion recognition, which has recently attracted much attention. The fusion approaches have improved accuracy by combining various types of information. This research proposes a fusion network with self-attention and local attention mechanisms. It uses a multi-layer perceptron network. The network extracts distinguishing characteristics from facial images using pre-trained models on RAF-DB dataset. We outperform the other fusion methods on RAD-DB dataset with impressive results.

1. Introduction

Facial expression recognition (FER) is essential to 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.

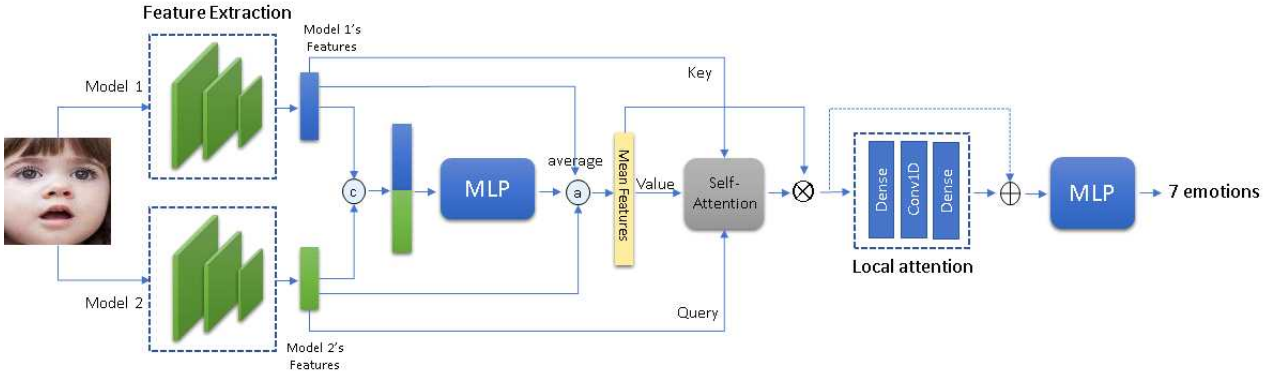
Attention mechanisms [1] have been frequently employed for problems including emotion classification in recent years and have produced better results. They focuses specific points on the face and comprehend the traits of each emotion to recognize the facial emotions. Besides, fusion approaches [2] are utilized as much as possible in the research, increasing the model's effectiveness.

This study proposed an attention fusion to produce significant features from two pre-trained models. We conducted the experiments on RAF-DB [3] dataset, compared to other fusion techniques and got excellent results.

2. Proposed Method

Our proposed method addresses the problem of emotion recognition from fusing two models for maximizing each model's benefits and reduce its drawbacks to enhance the performance.

Our approach involves creating an attention-based network that includes mechanisms for self-attention and local attention shown in Fig. 1. We employ a fusion method to combine the final features of two emotion recognition models that have already been trained. This combination aims to minimize each model's weaknesses while maximizing its strengths. We first concatenate the two features and pass them through a Multi-layer Perceptron (MLP) to generate a new feature the same size as the input sizes. We average the features before passing them through the self-attention and local attention



(그림 1) Local and Global Attention Fusion Network

blocks and then back and forth through a wholly connected network to classify emotions.

Self-attention (SA) [2]: It aims to exploit the contextual relationship between query Q, key K, and value V. In here, key and query (Q, K) elements are identified from the features of two pre-trained models. Before passing through a soft-max function, these features are multiplied by a dot product, then divided by the batch. Features from the MLP merged with two input features through averaging play the role of value (V). The overall equation is defined below:

$$SA(Q,K,V)=\left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V\right)\otimes V$$

where K, Q, V are key, query, and value, respectively. $\otimes$ is the element-wise multiplication.

Local Attention: It helps the model learn and comprehend information more effectively from a local spatial context and removes noises with two fully connected layers and a Conv1D layer. Besides, the residual connection is used to prevent the loss of unintended input features.

3. Experiments and Results

Experimental Setup: Our study used the basic set of RAF-DB dataset [3] to conduct experiments. The RAF-DB dataset contains 29672 real-world images that have been divided into two categories: basic emotions and compound emotions. Furthermore, the author has not yet published around 10,000 images that do not belong to the above two emotions. We only focus on basic emotion includes 15339 images, as well as a training and testing set. We trained conventional models such as VGG11 [4]. VGG13 [4], Resnet18 [5]. Resnet34 [5] to extract features to perform fusion approaches. The setting of the training configuration was 50 epochs, batch size 48, and learning rate 0.0001. The augmentation transformations such as flip, rotation in the range [-30,30], and remove saturation were used to avoid overfitting.

Results and Discussion: Firstly, we trained VGG11, VGG13, ResNet18, and ResNet34 models using Image-Net weight. The performance results were shown in Table 1.

<표 1> Performance of pretrained models

Model	ResNet18	ResNet34	VGG11	VGG13
Accuracy	85.3%	85.5%	85.23%	85.2%

After that, we applied fusion techniques [2] late fusion, early fusion, and joint-late fusion from extraction features of two pre-trained models. ResNet18 and ResNet34 models obtained the highest results for late fusion and early fusion, respectively, 86.35% and 86.66%, shown in Table 2. VGG13 and ResNet34 had the best results for joint fusion 86.63%.

<표 2> Comparision to fusion approaches

Fusion	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	86.40%
Our method	ResNet18	ResNet34	90.95%
	VGG13	ResNet34	**90.92%**

Results from our proposed approach are about 4% better than those from fusion approaches. In addition, the model is about 0.4 lower than the initial 90.48% when local attention is not used.

4. Conclusion

This research aims to achieve high efficiency in FER using a fusion network that combines attention mechanisms. The model uses a multi-layer perceptron combined with self-attention to filter out essential traits from the two models. The local attention mechanism helps verify features before passing them on to the classifier to classify the seven feelings. The proposed approach's higher accuracy compared to fusion approaches has been proven on the RAF-DB dataset. Using local attention enhances the feature selection process, and the model's accuracy is increased. Future research may focus on how effectively the suggested approach works with other datasets and related tasks.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A3A04036408) and also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub).

References

[1] Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M. and Hu, S.M., "Attention Mechanisms in Computer Vision: A Survey," Computational Visual Media, vol. 8, no. 3, pp. 331 - 368, Sep. 2022.

[2] Gadzicki, K., Khamsehashari, R. and Zetzsche, C., "Early vs Late Fusion in Multimodal Convolutional Neural Networks," in 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pp. 1 - 6, Jul. 2020.

[3] Li, S., Deng, W. and Du, J., "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, pp. 2584 - 2593, Jul. 2017.

[4] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," in 3rd International Conference on Learning Representations (ICLR), vol. 3, no. 1, pp. 1 - 14, Sep. 2015.

[5] He, K., Zhang, X., Ren, S. and Sun, J., "Deep Residual Learning for Image Recognition," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770 - 778, 2016.