Optimization in Facial Expression Recognition Based on CNN Combined with SE Modules

Xuanyu Zhang

Information and Computing Science, Shanghai Polytechnic University, Shanghai, 201209, China
2021120154@stu.sspu.edu.cn

Abstract. Facial expression recognition has emerged as a pivotal aspect of human-computer interaction and psychological research, drawing extensive attention in computer vision. The essay aims to improve the facial expression recognition performance of Convolutional Neural Networks (CNN) under different imaging conditions by combining attention mechanisms. In terms of data preparation, the FER-2013 dataset from Kaggle was used, which includes grayscale facial images with 48 x 48 pixels. By data augmentation and normalization, the diversity of the data is increased, and the robustness of the model is improved through random horizontal flipping, brightness and contrast adjustment, and the introduction of Gaussian noise. In terms of model architecture, a network structure similar to VGG is adopted, and a Squeeze and Excitation (SE) module is introduced after each convolutional layer, dynamically adjusting the importance of each channel through global average pooling and fully connected layers. The experimental results indicate that incorporating the attention mechanism reduces the model's loss across the training, validation, and test sets, while significantly enhancing its accuracy. These results demonstrate the effectiveness of attention mechanisms in facial expression recognition tasks. Overall, this study significantly improved the performance and robustness of CNN in facial expression recognition tasks by introducing attention mechanisms, demonstrating its superiority under complex imaging conditions.

Keywords: Facial expression recognition, convolutional neural networks, attention mechanisms

1 Introduction

Now in the realm of computer vision, recognizing facial expression stands as a critical component for applications ranging from human-computer interaction to psychological studies. Facial emotion recognition, as a specific application of facial recognition, aims to detect and classify different emotional states by analyzing facial expressions. This technology has gained widespread attention because it has been discovered that nearly 55 percent of the emotions in communication are delivered by facial expression [1]. Nowadays, many studies have demonstrated the potential of facial emotion recognition

© The Author(s) 2024
Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_99

technology in fields such as finance, healthcare, education, and human-computer interaction. As far as the education sector is concerned [2], facial emotion recognition can serve as a valuable observation tool to indicate student emotions and learning outcomes, thus providing personalized teaching methods and feedback mechanisms.

The main goal is to investigate how to enable machines to recognize minor changes in facial features, such as eyebrow movements and eye dilation, and facial expression recognition systems can accurately distinguish emotions such as happiness, anger, sadness, and surprise [3]. Although significant progress has been made in understanding and applying models [4], using Convolutional Neural Network (CNN) to accurately recognize facial expressions in different scenes such as lighting and exposure often poses challenges, which significantly impacting the model's accuracy.

As mentioned above, the adaptability of CNN to different imaging conditions remains unclear [5], especially in photos with insufficient lighting or overexposure, where key facial details may be obscured. The existing solutions mainly focus on enhancing model architecture or training strategies, without specifically addressing these common, real-world changes in image quality. This limitation greatly affects the effectiveness of facial expression recognition systems.

To further enhance the accuracy and interpretability of the proposed model, this study combined attention block with the model, which is to aim at improving feature extraction, enabling CNN to more effectively focus on prominent facial features crucial for accurate expression recognition. The main function of this method is to focus limited attention on key information. It can be roughly understood as using certain networks to compute weights, which are then applied to the feature map to modify it and obtain a feature map with enhanced attention.

A key aspect of this study is the introduction of attention mechanisms in the model, which allows the model to focus more on prominent facial features that are crucial for expression recognition, which has been demonstrated in other studies [6]. Through this method, the model can extract features more effectively and ignore or reduce the interference of noise. The main function of this attention mechanism is to focus limited attention on key information, thereby quickly obtaining the most effective information. The core logic is to shift from focusing on all content to focusing on key points [7]. The advantage of this approach is that the model can better focus on key information, thereby achieving better results when dealing with specific tasks. While this approach might lead to a higher number of parameters and increased model complexity, it can concentrate attention in key areas, thereby improving accuracy. In dealing with noise, attention mechanism reduces sensitivity to noise during the processing by making the model focus more on key information in the image. Specifically, when attention mechanisms are applied to a model, the model can learn to focus more attention on areas of significant importance for expression recognition in the image, while reducing attention to irrelevant information such as noise.

2 Method

2.1 Dataset Preparation

The FER-2013 dataset used in this article is from Kaggle [8] and consists of several 48×48-pixel grayscale images of faces. The faces have been automatically aligned, ensuring they are centered and occupy a similar amount of space in each image. The dataset includes 28,709 images for training and 3,589 images for testing, categorized into seven classes: 'angry', 'disgust', 'fear', 'happy', 'neutral', 'sad', and 'surprise'. Some sample images are displayed in Fig. 1.

Fig. 1. Samples of the dataset.

In terms of data preprocessing, this study first performs data augmentation and normalization on the dataset, increasing data diversity through random horizontal flipping, brightness and contrast adjustments, and scaling pixel values within the range of [0,1]. Next, adjusting the size to the same size for easier model processing. Also creating DataLoader, it is to effectively load data and provide it to the model for training, validation and testing. In addition, in order to effectively load data and provide it to the model for training, validation, and testing, the method sets batch sizes and randomly shuffles the sample order within each training cycle to help the model learn better. By specifying the number of child processes, the data loading speed is accelerated. Meanwhile, loading data into the fixed memory of the GPU can accelerate data transfer. This method can efficiently provide data streams and provide necessary inputs for model training. To further enhance the model's suitability for noise, various forms of noise

were introduced, including the random addition of Gaussian noise. This series of preprocessing operations ensures data consistency and diversity, providing richer input information for the model. Finally, some sample images with added noise shown in Fig. 2 were presented to visually demonstrate the data processing effect.

Fig. 2. Transformed Images with Gaussian Noise

2.2 Attention-based CNN

Convolutional Neural Network is a type of deep learning model commonly employed for image processing tasks [9-12]. It is composed of convolutional and pooling layers. The convolutional layers are responsible for feature extraction from images, using filters to detect specific patterns. Each filter identifies particular characteristics or features within the images. The pooling layers perform downsampling to reduce the dimensionality of the feature maps, which helps maintain essential feature information while decreasing computational complexity and mitigating the risk of overfitting.

Attention mechanism shown in Fig. 3, as previously mentioned, is a technique used to enhance model performance, with the main idea of giving the model varying degrees of attention to different parts of the input. And in this article, it is to improve the model's accuracy and generalization capability. Squeeze and Excitation Networks (SENet) [13] is a classic attention mechanism model that dynamically adjusts the importance of each

channel by learning the relationships between channels, mainly including squeeze, excitation and scale operations. Squeeze is a global average pooling of the input feature layers, compressing the features of each channel into a global feature value to obtain a channel descriptor. This descriptor reflects the information of the channel on a global scale. Excitation is the process of learning the weights of each channel through a two-layer fully connected network after obtaining the channel descriptor. The function of these two fully connected networks is to capture the dependency relationships between channels and generate a weight value for each channel based on these dependencies. This weight value reflects the importance of the channel for the final output feature. Scale is achieved by multiplying the weight values of each channel by the corresponding channel of the original input feature layer, in order to reweight the features. This process can be seen as a calibration of the original features, allowing the network to focus more on important feature channels and suppress irrelevant feature channels.

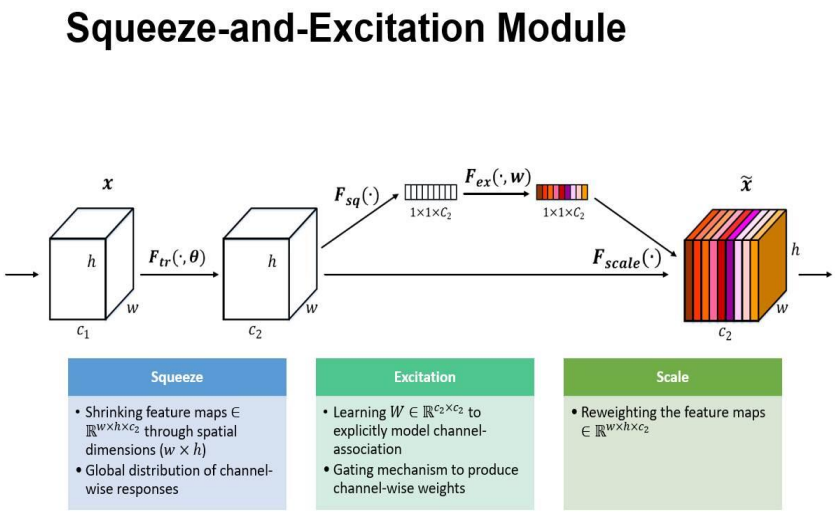

Fig. 3. Function of SEnet.

In the proposed attention-based CNN structure, attention mechanisms are introduced into certain layers of convolutional neural networks to enhance the model's attention to specific features. Combining the characteristics of CNN and attention mechanism, this structure can effectively extract key features from images and dynamically adjust the importance of features through attention mechanism, thereby improving model performance.

Specifically, the CNN architecture consists of multiple convolutional and pooling layers, each followed by batch normalization and ReLU activation functions. The initial layer employs a 3×3 convolution kernel, generating 32 feature channels from a single-channel input image. Subsequent layers also use 3×3 convolution kernels, with the number of feature channels increasing to 64 and then 128. Each convolutional layer is succeeded by a pooling layer, where 2×2 max pooling is applied to reduce the spatial dimensions of the feature maps, thus decreasing computational complexity.

Attention mechanisms are integrated into the model through SE Modules, placed after each convolutional layer. The SE module compresses the features of each channel into a scalar through global average pooling and generates importance weights for each channel using fully connected layers and activation functions. These weights are then applied to the feature representation of each channel through element-wise multiplication. This enables the model to dynamically adjust the channel weights based on global information, thereby enhancing or suppressing specific feature representations.

For the classification phase, the model employs fully-connected layers. The feature map is first flattened into a one-dimensional vector. This is followed by two fully connected layers, each containing 1,024 neurons and using ReLU activation functions and a Dropout mechanism. The final output is the prediction for seven categories. The integration of the attention mechanism enables the model to better focus on task-related features, thus improving classification performance.

2.3 Implementation Details

This experiment implemented an emotion recognition model based on CNN, using attention mechanism in SENet to enhance model performance. The model's architecture resembles the VGG network, comprising convolutional layers, pooling layers, and fully connected layers. To train the model, the Adam optimizer [14] and the cross-entropy loss function were utilized, and an early stopping strategy was employed to mitigate overfitting by monitoring the validation set's loss value. During training, the loss values and accuracies of both the training and validation sets were recorded for subsequent visualization and analysis. The model's training hyperparameters were set to 50 epochs with a batch size of 32. The entire model construction and training were conducted using the PyTorch framework. After training, the model was evaluated on the test set, and the loss values and accuracy on this set were calculated. Visual results of the training and validation metrics are provided to further analyze the model's performance.

3 Results and Discussion

3.1 The Performance of Models

In the training curves shown in Fig. 4, it was observed that there was no significant difference in the growth trend of accuracy and loss with or without attention mechanism. Therefore, the final results are presented in the form of a table for analysis.

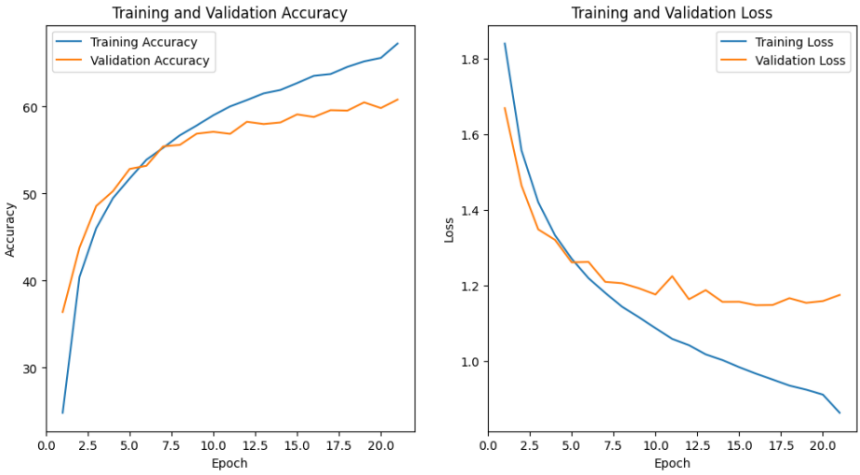

Fig. 4. The training curves.

Table 1 shows the changes in training and validation accuracy with and without attention mechanism.

Table 1. Loss and Accuracy

Dataset	Loss without attention	Accuracy without attention	Loss with attention	Accuracy with attention
Training	1.0126	61.56%	0.8622	67.24%
Validation	1.2765	55.63%	1.1741	60.80%
Test	1.1077	58.61%	1.0554	61.31%

The results in Table 1 show that when model is without attention mechanism, the training loss of the model is 1.0126, and the training accuracy is 61.56%; The validation loss is 1.2765, and the validation accuracy is 55.63%; The test loss is 1.1077, and the test accuracy is 58.61%. It can be seen that without attention mechanism, the training accuracy and validation accuracy of the model are relatively low, and the growth rate is slow during the training process. In contrast, when model is with attention mechanism, the training loss of the model is reduced to 0.8622, and the training accuracy is improved to 67.24%; The validation loss was reduced to 1.1741, and the validation accuracy increased to 60.80%; The testing loss was 1.0554, and the testing accuracy was improved to 61.31%. It can be seen that without attention mechanism, the training accuracy and validation accuracy of the model are relatively low, and the growth rate is slow during the training process.

3.2 Discussion

Comparing the results with and without attention mechanisms reveals significant improvements in the model's overall performance when attention mechanisms are employed. Specifically, the model's loss on the training, validation, and testing sets is reduced, and the accuracy is enhanced. This demonstrates that attention mechanisms enable the model to focus more effectively on important features, thereby enhancing its adaptability and robustness.

The findings indicate that attention mechanisms allow the model to dynamically adjust the significance of feature channels during feature processing, which improves the model's focus on key features. This adjustment not only boosts accuracy but also strengthens the model's performance and resilience. This not only improves the learning efficiency of the model, but also enhances its generalization ability on unseen data. It can be inferred that in models without attention mechanisms, although CNN performs well in feature extraction, there are still shortcomings in processing noise in images. CNN is often disturbed by local features, such as random noise or non-key features in images, which makes it easy for the model to focus on unimportant information during the learning process, thereby affecting overall performance.

Although the introduction of attention mechanism significantly improves model performance, there are still some limitations in this study. For example, this article fails to fully validate the scientific basis for the effectiveness of attention mechanisms. Specifically, although attention mechanisms seem to improve performance, there is no clear evidence that they focus on useful or important regions in the image. This study has not yet conducted sufficient visualization and scientific analysis to confirm that the performance improvement is indeed due to the correct application of attention mechanisms to key features. The improvement of the model may be accidental, or the areas emphasized by the attention mechanism may not be as meaningful as expected.

In summary, the attention mechanism significantly enhances the model's performance in emotion detection tasks by enhancing its ability to dynamically adjust features. Future work can further optimize attention mechanisms, such as adjusting the number of feature channels to explore more advanced combinations of attention modules, and introducing more diverse data augmentation techniques to further improve model performance and robustness.

4 Conclusion

The study significantly enhanced the performance of CNN in emotion detection tasks by incorporating attention mechanisms. A comparison of models with and without attention mechanisms reveals that attention mechanisms significantly improve the model's accuracy and robustness. Specifically, attention mechanisms dynamically adjust the importance of feature channels, enabling the model to concentrate more on key features, reducing sensitivity to noise and irrelevant features, and thus improving overall performance.

Experimental results demonstrate that introducing attention mechanisms reduces loss and improves accuracy across all datasets. These findings underscore the effectiveness of attention mechanisms in enhancing model performance. While attention mechanisms have substantially boosted performance, future research could focus on further optimizing these mechanisms and employing diverse data augmentation techniques to enhance model performance and generalization ability.

References

1. Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pp. 568-573. IEEE (2005).
2. Wang, W., Xu, K., Niu, H., et al.: Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. Complexity 2020, pp. 1-9 (2020).
3. Warnick, B. J., Davis, B. C., Allison, T. H., et al.: Express yourself: Facial expression of happiness, anger, fear, and sadness in funding pitches. Journal of Business Venturing 36(4), 106109 (2021).
4. Nonis, F., Barbiero, P., Cirrincione, G., et al.: Understanding abstraction in deep CNN: an application on facial emotion recognition. Progresses in Artificial Intelligence and Neural Systems, pp. 281-290 (2021).
5. Ferrante, E., Oktay, O., Glocker, B., et al.: On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains. In Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9, pp. 294-302. Springer International Publishing (2018).
6. Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., Guo, L.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 103323 (2022).
7. Zhu, H., Xie, C., Fei, Y., et al.: Attention mechanisms in CNN-based single image super-resolution: A brief review and a new perspective. Electronics 10(10), 1187 (2021).
8. Kaggle: FER2013. Available at https://www.kaggle.com/datasets/msambare/fer2013 (2013).
9. Qiu, Y., Chen, H., Dong, X., Lin, Z., Liao, I. Y., Tistarelli, M., Jin, Z.: Ifvit: Interpretable fixed-length representation for fingerprint matching via vision transformer. arXiv preprint arXiv:2404.08237 (2024).
10. Liu, Y., Bao, Y.: Review on automated condition assessment of pipelines with machine learning. Advanced Engineering Informatics 53, 101687 (2022).
11. Liu, Y., Bao, Y.: Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. Measurement 220, 113418 (2023).
12. Ye, X., Luo, K., Wang, H., Zhao, Y., Zhang, J., Liu, A.: An advanced AI-based lightweight two-stage underwater structural damage detection model. Advanced Engineering Informatics 62, 102553 (2024).
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141 (2018).
14. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.