Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/22/20231157

Convolutional neural network combined with the attention mechanism for facial emotion recognition

Xiguo Luo

The Department of Natural Science, Durham University, DH13LE, United Kingdom

xiguo.luo@durham.ac.uk

Abstract. Facial Emotion Recognition (FER) holds great importance in the fields of computer vision and machine learning. In this study, the aim is to improve the accuracy of facial expression recognition by incorporating attention mechanisms into Convolutional Neural Networks (CNN) with FER2013 dataset, which consists of grayscale images categorized into seven expressions. The combination of proposed CNN architecture and attention mechanisms is thoroughly elucidated, emphasizing the operations and interactions of their components. Additionally, the effectiveness of the new model is evaluated through experiments, comparing its performance with existing approaches in terms of accuracy. Besides, the results demonstrate that the CNN architecture with attention mechanisms outperforms the original CNN by achieving an improved accuracy rate of 69.07%, which is higher than 68.04% accuracy rate of original CNN. Moreover, the study further discusses the confusion matrix analysis, revealing the challenges faced in recognizing specific emotions due to limited training data and vague facial features. In the future, this study suggests addressing these limitations through data augmentation and to reduce the gap between training and testing accuracy. Overall, this research highlights the potential of attention mechanisms in enhancing facial expression recognition systems, paving the way for advanced applications in various domains.

Keywords: convolutional neural network, deep learning, FER.

1. Introduction

FER is a perspective research topic in computer vision and machine learning. In the 1970s, the definition of six human facial expression, which are happiness, anger, surprise, fear, disgust and sadness, was provided by Ekman and Friesen with a number of experiments [1]. Afterwards, the neural expression is also considered as a category of expressions. Nowadays, facial expression recognition has broadened research prospects has opened up new avenues of research various domains, including human-computer interaction, and affective computing, emotion analysis, intelligent security and entertainment etc. According to the report of Mordor Intelligence in 2022, the Emotion Detection and Recognition (EDR) market was valued at $19.87 million in 2020. It is projected to reach USD 52.86 million by 2026, at a CAGR of 18.01% during the forecast period (2021-2026) [2]. This indicates a significant market potential for the application of facial expression recognition within the next three years.

From previous studies, Convolutional Neural Networks (CNN) was successfully used to analyze and classify images, including those in the FER2013 dataset. Several influential works have shaped

DOI: 10.54254/2755-2721/22/20231157

the development and advancement of CNN architectures. In 2012, AlexNet introduced a new concept of using deep architecture of CNN and applied their ability to learn hierarchical representations, achieving a breakthrough in the ImageNet challenge [3]. In 2014, the VGGNet architecture introduced by Simonyan and Zisserman emphasized the importance of increasing network depth and using smaller filters to capture more detailed features [4], resulting in improved performance. Moreover, the InceptionNet architecture presented Inception module concept to capture features from multi-scale by combining filters of different sizes within a layer [5]. However, the previous researches in FER have predominantly concentrated on the design of convolutional layers without considering the weight allocation of various learned feature representations. To further enhance the performance of CNN for FER, attention mechanisms have been introduced. With the attention mechanism, network can selectively focus on relevant facial regions to capture important discriminative features. Based on the foundation laid by previous studies, the core difference of this research lies in the incorporation of attention mechanisms. By leveraging attention mechanisms into each convolutional layer in CNN architecture, the aim to achieve higher accuracy in facial expression recognition on the FER2013 dataset.

The study will present a comprehensive overview of the proposed CNN architecture with attention mechanisms for facial expression recognition. The architecture will consist of convolutional layers, pooling layers, and attention modules with explanations of the operations and interactions of these components, demonstrating how attention mechanisms improve the model's accuracy rate. To evaluate the effectiveness, the performance of related experiments will be conducted on the FER2013 dataset with comparison of the proposed architecture with traditional CNN, considering matrix such as accuracy. Through these experiments, the purpose is to demonstrate the superiority of the CNN integrated with the attention mechanism in accurately classifying facial expressions. The primary contribution of the research is the development of a CNN architecture with attention mechanisms specifically tailored for facial expression recognition on the FER2013 dataset. The result demonstrated that the accuracy of CNN will be improved with the attention mechanism compared to the original CNN.

2. Method

2.1. Dataset description and preprocessing

The FER2013 dataset is a widely recognized as a prominent publicly available dataset employed for facial expression recognition with 28709 training images and 7178 testing images. It was released as part of the challenge of FER at the International Conference on Machine Learning (ICML) in 2013 [6]. The images are a grayscale with dimensions of 48×48 and can be classified into seven expressions.

This study uses the ImageDataGenerator class from the Keras library, which offers a convenient approach to preprocess image data with a number of data augmentation options [7]. There are three data augmentation techniques used in Image. Firstly, the rescale technique normalizes the pixel values between 0 and 1 with the parameter set to 1./255. This normalization step plays a crucial role in ensuring stability and efficiency of the training process. Additionally, two data augmentation techniques, namely width and height shift, are applied to the training images. These techniques randomly shift the images in a horizontal and vertical direction, respectively, by a fraction of their total width or height. By applying these shifts, the model becomes more robust to variations in the position of facial features, enhancing its ability to generalize to different face orientations and alignments.

The second data technique used is "flow_from_directory" method to load both training and test images directory. Firstly, the target_size parameter is put to (48, 48) to ensure consistent input dimensions for the model by resizing the image to a resolution of 48x48 pixels. Secondly, the parameter of "color_mode" is set to 'grayscale', presenting that the images will be converted to grayscale during the loading process to reduce the computational complexity and memory requirements for training the model, as grayscale images contain only one-color channel instead of

DOI: 10.54254/2755-2721/22/20231157

three (red, green, and blue) in RGB images. Thirdly, the parameter of "class_mode" is set to 'categorical', showing that the class labels are represented as one-hot encoded vectors, which allow the model to predict the probabilities of each emotion category independently. Lastly, the parameter of "batch_size" is set to 64, specifying the number of input images to be distributed to each mini-batch during training. With the use of mini-batches, it helps to accelerate the training process by allowing the model to update its parameters more frequently.

2.2. CNN combined with attention module-based emotion recognition model

2.2.1. Introduction of CNN. CNN are composed of three components: the convolutional layer, pooling layer, and fully connected layer. The convolution layer has the capability to automatically adapt and extract relevant features [8]. Besides, the pooling layer in CNN performs downsampling by reducing the spatial size of feature maps to retain important features [9]. Additionally, the fully connected layer follows the flatten layer to provide global spatial information and helps in making predictions [10].

2.2.2. Introduction of SENet. Squeeze-and-Excitation Network (SENet) is a deep learning architecture that incorporates an attention mechanism called the SE block shown in Figure 1, which enhances the representation power of CNN by adaptively recalibrating feature responses [11]. Besides, the SE block learns to selectively emphasize discriminative features and suppress less relevant ones, improving the discriminative ability of the model. By explicitly modeling the interdependencies among the channels, SENet enables CNN to focus on more informative features and improve higher performance in FER.

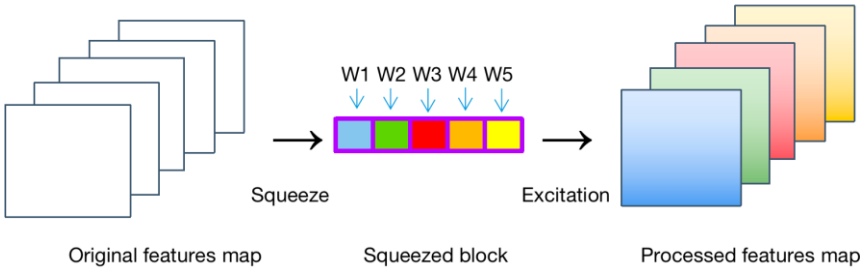


Figure 1. The structure of SE block.

2.2.3. The structure of CNN with SE Block. The structure of CNN follows one of the most accurate models in Kaggle presented by Mohamed Chahed with three convolutional blocks and max-pooling layers [12]. Additionally, a fully-connected layer and a softmax layer follow the flatten layer. In each convolution block, there are two convolutional layers with 'Relu' activation to capture the features of input data with max-pooling layer. Then, the flatten layer resizes the data to suit the fully connected layer. Lastly, the softmax layer with 7 types of emotions facilitates the model to classify the task. Different from the original model, the CNN with SE Block shown in Figure 2 before max-pooling layer, it reorganizes the weights of feature maps to perform better by attention on important features.

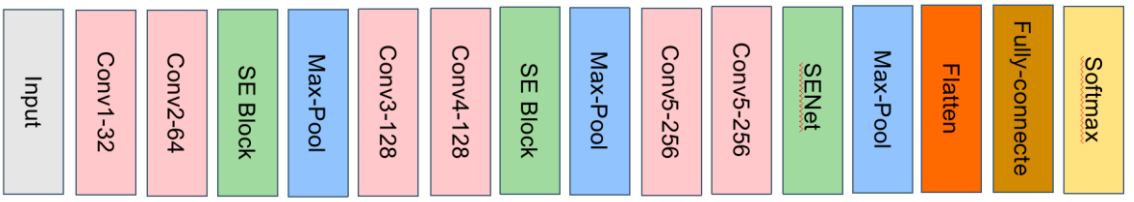


Figure 2. The structure of CNN with SE blocks.

2.3. Implementation details

The implementation of models in this study is based on the Tensorflow. Prior to the training of the

DOI: 10.54254/2755-2721/22/20231157

model, this study is divided both training data and test date by set batch size to help the model learn better representations and improve its generalization capabilities. Additionally, there are a number of techniques used during the training process of model, which are EarlyStopping, ModelCheckPoint and ReduceLROnPlateau. For EarlyStopping, it monitors the accuracy of validation and stops training if no improvement is observed for a certain number of epochs defined by the patience parameter [13]. In this study, the patience is set to 16 to reach the limit of this model. For ModelCheckPoint, it saves the weights of model only when validation accuracy improves, ensuring that the best-performing model is saved. For ReduceLROnPlateau, it reduces the learning rate when the validation accuracy stops. With this adjustment of the learning rate, it helps fine-tune the model's progress and overcomes accuracy plateaus [14]. During compiling the model, the loss function with categorical cross-entropy, Adam optimizer with a learning rate of 0.002 and the accuracy matrix are used to reach the limit of performance, respectively. Besides, there are 100 epochs to train the model.

3. Results and discussion

3.1. The performance of the model

Table 1 presents the comprehensive results divided into four categories: training loss, training accuracy, testing loss, and testing accuracy. Both models have higher training accuracy than testing accuracy and lower training loss than testing loss. After applying the attention mechanism into CNN, the performance of CNN with SE Blocks outperformed than the original. The accuracy of the modified model demonstrates an improvement of 1.03%, reaching 68.04% compared to the baseline. However, the testing loss of CNN with SE Blocks is higher than the CNN. Overall, the incorporation of attention mechanisms yields enhanced performance for CNN in this study.

Table 1. The performance of the various models.

	Training Loss	Training Accuracy	Testing Loss	Testing Accuracy
Original CNN	0.3767	86.21%	1.1134	68.01%
CNN with SE Blocks	0.2913	89.45%	1.2447	69.04%

3.2. Discussion

In this study, the confusion matrices of the models on FER-2013 are calculated shown in Figure 3. Focusing on the diagonal of two matrices, the prediction of the “happy” label is the best, reaching more than 4 hundred correct data, while the “disgust” label is the worst. The possible reasons are listed as follow: 1) Few training data: Of all kinds of the labels, the “disgust” label has the fewest training data – 93 horizontally added. However, the “happy” label has 1, 773 training data, nearly 17 times compared with the “disgust” label. Thus, the few training samples lead to low accuracy. 2) Vague facial feature: The disgust facial expression has little unique feature that is different from other emotions. A “disgust” labeled image is extremely likely to be identified as similar emotions, such as “angry” and “fear” according to the confusion matrix.

In light of the limited availability of data pertaining to the emotion of disgust, it would be advantageous to employ data augmentation techniques as a means to address this shortcoming in future research endeavors focused on broadening the scope of facial expression recognition. By utilizing data augmentation, the augmentation of the existing dataset can effectively alleviate the scarcity of disgust-related data, thereby enabling the expansion of the current boundaries within the realm of FER [15].

DOI: 10.54254/2755-2721/22/20231157

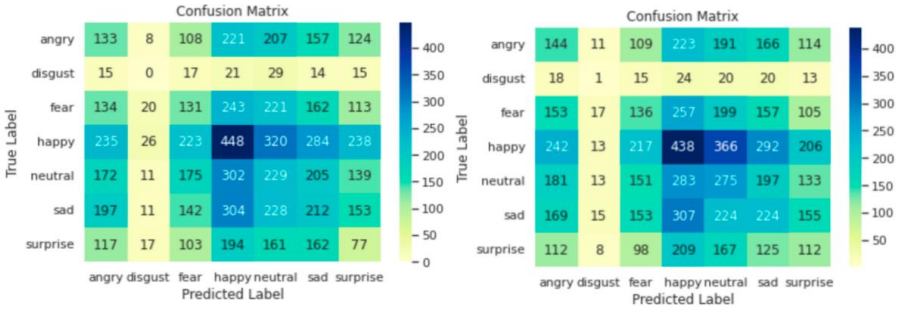
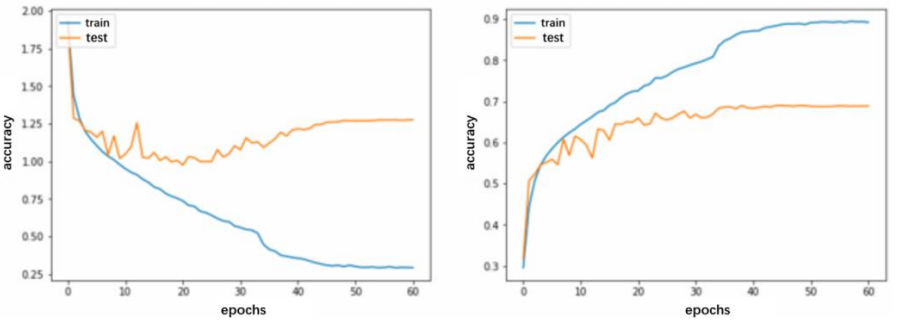


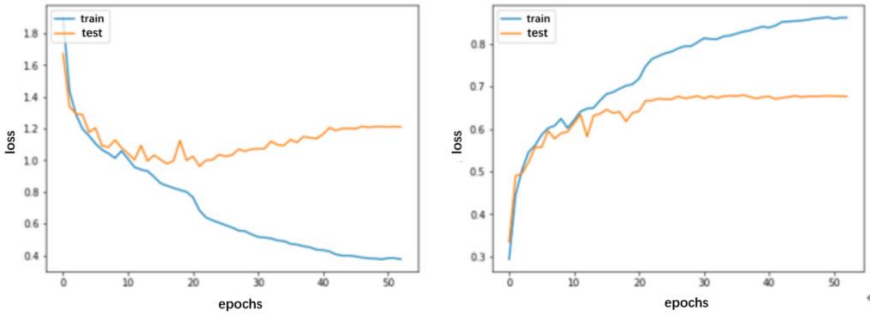
Figure 3. Confusion matrix of CNN and CNN with SE blocks.

The variations of the training loss, test loss and the accuracy changing on each epoch are shown in Figure 4 Based on the FER-2013 dataset, the loss of the self-build model is decreasing with the increase of the epochs, while the accuracy is increasing. When the model runs to approximately 20 epochs, the curve becomes steady gradually. Around the 60 epochs, the early stopping mechanism comes into effect, halting further model fitting as the validation accuracy fails to surpass the best achieved accuracy.

Compared with the CNN, the CNN with SE Blocks model has reached the similar effect and learning speed, and two final accuracies nearly reach 70%. However, the gap between training accuracy and testing accuracy has increased more and more. Especially, CNN with SE Block has further improvement in around 36 epochs with the decreased learning rate. Thus, SE Block can better perform the accuracy of facial expression recognition. Additionally, the method to prevent over-fitting problem should be used to decrease the gap between training accuracy and testing accuracy to get better performance. Furthermore, the new structure with SE block after each convolutional layer has a significant improvement to CNN, which has been demonstrated in many studies [16-18]. As a result, there is still a space for improvement in the convolution block.



(a)



(b)

Figure 4. Epochs-accuracy curve related to the (a) CNN (b) CNN with SE blocks.

DOI: 10.54254/2755-2721/22/20231157

4. Conclusion

In conclusion, the integration of attention mechanisms into the convolutional layer of CNN presents a creative and promising approach for enhancing accuracy in facial expression recognition tasks. With the combination of SE block into traditional CNN's architectures, the model can focus on the most informative regions or features within an image to capture and emphasize relevant information. Additionally, not only the performance in the field of facial expression recognition, but also of the whole field of CNN can be improved by the creative structure of SE block. Compared to original CNN, the CNN with SE block achieved a higher accuracy of both training accuracy and testing accuracy during processing and result and controlled the focus region by modifying the loss function. However, the limitation of the new model is the control the focus region by modifying the loss function. In the future work, the performance of the new model can be improved by setting parameter with experience.

References

[1] Ekman P and Friesen W 1978 The facial action coding system: a technique for the measurement of facial movement (Santa Clara: Consulting Psychologists Press)
[2] Mordor Intelligence 2023 Emotion Detection and Recognition (EDR) Market - Growth, Trends, COVID-19 Impact and Forecast (2023 - 2028) https://www.mordorintelligence.com/zh-CN/industry-reports/emotion-detection-and-recognition-edr-market
[3] Krizhevsky A and Sutskever I and Hinton G E 2012 ImageNet Classification with Deep Convolutional Neural Networks (In Advances in Neural Information Processing Systems) pp 1097-1105
[4] Simonyan K and Zisserman A 2014 Very Deep Convolutional Networks for Large-Scale Image Recognition (arXiv preprint arXiv) 1409-1556
[5] Szegedy C and Liu W and Jia Y and Sermanet P and Reed S and Anguelov D and Rabinovich A 2015 Going Deeper with ConvolutionsIn Proceedings of the IEEE conference on computer vision and pattern recognition pp 1-9
[6] Carrier P L and Courville A 2013 Challenges in Representation Learning: A Report on Three Machine Learning Contests (Neural Information Processing Systems) Workshop on Challenges in Representation Learning
[7] Chollet F 2018 Deep learning with Python Manning Publications
[8] Pramerdorfer C and Kampel M 2016 Facial expression recognition using deep learning: A survey IEEE
[9] LeCun Y and Bengio Y and Hinton G 2015 Deep learning (Nature) 521 (7553) pp 436-444
[10] Goodfellow I and Bengio Y and Courville A 2016 Deep learning (MIT press)
[11] Hu J and Shen L and Sun G 2018 Squeeze-and-Excitation Networks (CVPR) pp 7132-7141 (IEEE)
[12] Chahed M 2021 Human Emotion Detection https://www.kaggle.com/code/mohamedchahed/human-emotion-detection
[13] Prechelt L 1998 Early stopping-but when? (Neural Networks: Tricks of the Trade) pp 55-69 (Springer)
[14] Smith L N 2015 Cyclical learning rates for training neural networks (WACV) pp 464-472 (IEEE)
[15] Raschka S and Mirjalili V 2020 Python Machine Learning (3rd ed) (Packt Publishing) Chapter 6
[16] Yao L Xiao X Cao R et al 2020 Three stream 3D CNN with SE block for micro-expression recognition 2020 International Conference on Computer Engineering and Application (ICCEA) IEEE 2020 pp 439-443
[17] Yu Q Wang J Jin Z et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72: 103323

DOI: 10.54254/2755-2721/22/20231157

[18] Rouvier M Bousquet P M 2021 Studying squeeze-and-excitation used in CNN for speaker verification 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) IEEE pp 1110-1115