Layer-wise Interpretability Investigation of Facial Expression Recognition Models Based on Grad-CAM

Siyuan Yao

[1] Computer Science and Technology, North China University of Technology, Beijing, 100144, China
Email:21101020110@mail.ncut.edu.cn

Abstract. For a long time, artificial intelligence has faced the challenge of interpretability, with the black-box problem persistently troubling researchers. Although there have been studies using Gradient-weighted Class Activation Map (Grad-CAM) for interpretability in the field of facial expression recognition, these studies often lack attention to the impact of each layer of the model on interpretability. In this study, different models are constructed, and Grad-CAM is used to explore the impact of different types of layers on model interpretability, filling the gap left by previous work. To be more specifically, this study constructed various Convolutional Neural Networks (CNN) models, including a baseline model, three models with modified convolutional layers, and three models with modified pooling layers. For comparative experiments, all six modify models were modified from the baseline model. All these models were trained using the FER-2013 dataset. Before training, the dataset underwent image preprocess and augmentation to prevent overfitting. After training these models, Grad-MAPs are generated based on the same test images. Experimental results show that different layers significantly impact model interpretability: convolutional layers affect the size of hotspot regions, while pooling layers influence the discreteness of these regions.

Keywords: Convolutional Neural Networks, Facial Expression Recognition, Gradient-weighted Class Activation Mapping.

1 Introduction

Humans express their emotions to others through various means such as language, gestures, body language, and facial expressions. Studies indicate that emotional information accounts for more than half of people's social behaviors, and facial expressions play a crucial role in expressing emotions during social interactions [1]. As a result, Facial Expression Recognition (FER) technology has significant potential for development and application in areas such as detecting fatigue and assessing mental health. With the rise of Artificial Intelligence (AI) technology in recent years, FER has been gradually implemented using Convolutional Neural Networks (CNN). By extracting features from human facial images, CNN can classify the expression in different emotions effectively.

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_102

Although facial expression has made significant progress with the help of AI, there exists a critical issue in these AI models known as the "black box" problem [2] although many studies have been carried out in different fields [3-6]. This refers to the opacity of AI systems; where it is often unclear how and why certain decisions are made within these models. This kind of lack of transparency will cause serious challenges, particularly when AI facial expression classification is used in sensitive domains such as healthcare or autonomous driving, where understanding the basis for decisions is crucial.

Researchers have always been striving to enhance the accuracy of model recognition, and recent studies have used Gradient-weighted Class Activation Mapping (Grad-CAM) to understand model decisions through focus areas [7, 8]. However, those researches have largely neglected a detailed layer-by-layer analysis of CNNs. Each layer of a CNN processes the input image differently, starting from simple feature recognition like edges and textures to more complex features like facial components in higher layers. Understanding the transformations and role of each specific layer in processing and classifying emotions could provide deeper insights into the workings of FER systems. This granular analysis is essential not only for making AI models more interpretable but also for refining them. Enhanced transparency can lead to better trust and reliability, thus facilitating broader acceptance and application of this technology.

This paper explores the intricacies of facial expression recognition by constructing and analyzing several custom-designed CNN models. To delve into the underlying mechanisms of these models and try to explain the prevalent "black box" issue, by employing Grad-CAM as the primary investigative tool. Grad-CAM enhances the interpretation of CNN decisions by providing visual explanations of the network's internal workings, highlighting important regions in the image that contribute to a specific classification decision. By comparing the activation mappings across different models with Gard-CAM to evaluate how variations in architecture influence the interpretability and performance of the system. This paper will try to systematically dissect each layer and parameter of these custom-designed CNN models to understand how each contributes to the overall task of emotion recognition. Not only can identify the most salient features at each layer but can also reveal how these features integrate to form the final decision output. This layer-by-layer analysis will potentially allow the study to pinpoint redundancies or inefficiencies within the model architectures, offering opportunities for optimization and refinement.

2 Method

2.1 Data Preparation

Dataset introduction. This study uses the FER-2013 dataset collected from Kaggle [9]. The images in this data set are all consist of 48x48 pixel grayscale images of human faces. The faces have been placed in the center of the image and occupy about the same amount of space in each image. All images correspond to one of the following labels: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral

Dataset Splitting. This study divided the FER-2013 into three datasets: training, validation, and testing, to effectively train, validate, and ultimately test the generalization ability of the model. The 28,709 images provided by FER-2013 were split into 80% for the training set and 20% for the validation set.

Data Pre-process and Augmentation. This study has applied data augmentation techniques to enhance the model's generalization ability to new data. This is particularly beneficial in situations where the available training data is limited. By introducing minor variations, data augmentation effectively simulates new data and thus significantly expands the diversity of training samples. This study has used the following methods.

Pixel Value Rescaling. This study has transformed pixel values linearly, from the range of 0-255 integers to 0-1 floats. This scaling is a standard preprocessing step aimed at easing the handling of input values by neural networks, facilitating smaller-range processing.

Random Rotation. Images have been rotated randomly with a maximum 10 degrees. This randomness helps increase the model's robustness to variations in head orientation.

Random Shifting. Images have been applied random shifts to the images horizontally and vertically, with maximum shifts of 10% and 5% of the image's width and height respectively. These transformations help the model maintain.

Horizontal Flipping. Images have been randomly flipped horizontally. This is beneficial for facial expression recognition as expressions typically retain their significance even when mirrored.

2.2 Self-designed CNN

CNNs are a type of deep learning model specifically designed for processing image and video data [10, 11]. Inspired by the neural connectivity patterns in biological visual systems, CNNs effectively capture spatial and local features in images. CNNs offer advantages such as automatic feature extraction, parameter sharing, and local connectivity. CNNs are widely used in various applications, including image classification, object detection, image segmentation, facial recognition, and autonomous driving, significantly improving the performance of computer vision tasks.

This study employs several custom-built CNN models, with a basic CNN model used to provide a baseline, the structure of the reference model is provided in Table 1. The basic model includes 5 conv-layers, a flatten-layer, and 5 dense-layer. This model also used batch-normalization and dropout to prevent overfitting.

Table 1. The architecture of the basic CNN model

Type	Output Shape	Param
Conv2D	(None, 46, 46, 32)	896
BatchNormalization	(None, 46, 46, 32)	128
Conv2D	(None, 46, 46, 64)	18,496
MaxPooling2D	(None, 23, 23, 64)	0

Conv2D	(None, 21, 21, 64)	36,928
BatchNormalization	(None, 21, 21, 64)	256
MaxPooling2D	(None, 11, 11, 64)	0
Conv2D	(None, 9, 9, 128)	73,856
Conv2D	(None, 9, 9, 256)	295,168
BatchNormalization	(None, 9, 9, 256)	1,024
MaxPooling2D	(None, 5, 5, 256)	0
Dropout	(None, 5, 5, 256)	0
Flatten	(None, 6400)	0
Dense	(None, 512)	3,277,312
Dense	(None, 256)	131,328
BatchNormalization	(None, 256)	1024
Dense	(None, 128)	32,896
Dense	(None, 64)	8,256
BatchNormalization	(None, 64)	256
Dropout	(None, 64)	0
Dense	(None, 32)	2080
BatchNormalization	(None, 32)	128
Dropout	(None, 32)	0
Dense	(None, 7)	231

2.3 Gradient-weighted Class Activation Mapping

Grad-CAM is a technique for visualizing the decision-making process of convolutional neural networks, enhancing understanding of how models make predictions. The fundamental principle of Grad-CAM is to utilize gradient information specific to a class of interest to highlight the most important regions in an image for predicting that class.

Specifically, Grad-CAM calculates the gradients of the last convolutional layer's feature maps for a target class output. These gradients are aggregated through global average pooling and applied as weights to the same convolutional layer's feature maps. The weights for each feature map effectively reflect their importance in predicting the target class. By performing a linear combination of these weighted feature maps, a coarse class activation map (CAM) is obtained, which spatially highlights the regions most crucial for the model's classification decision [12].

By overlaying this activation map on the original image, Grad-CAM provides an intuitive visual explanation, showing which areas are critical when the model makes a prediction. This technique is significant for understanding how complex convolutional networks operate in practical applications, especially when explaining predictions, reviewing models, and analyzing errors is necessary.

2.4 Implementation Details

This study starts with constructing several CNN models using TensorFlow, comprising a baseline model and three experimental variants to explore different architectural nuances. Each model was trained on the same dataset to ensure uniformity in training conditions and data exposure. The training parameters are 64 batch size and 45 epochs for each model. Once training is finished, A Grad-CAM image will be generated. All models use the same images to generate Grad-CAM images. The Adam optimizer is employed for its efficiency in handling sparse gradients and its adaptive learning rate capabilities, crucial for the model's fast convergence. The initial learning rate is set to 0.001.

Categorical cross-entropy is used as it is well-suited for multi-class classification problems. This loss function compares the distribution of the predictions with the true distribution of the labels, effectively guiding the training process toward accurate emotion recognition. The early stop function is used as callbacks to prevent the overfitting problem and to optimize training time. Specifically, the validation accuracy (val_accuracy) is set to be a crucial metric to gauge the performance of our models on unseen data. The patience parameter is 8, which means that if the validation accuracy does not improve in eight consecutive epochs, the training process will automatically stop. This approach helps in ensuring that all the models achieve the best generalization performance without wasting computational resources on training epochs that do not yield significant improvements.

3 Result and Discussion

3.1 The Performance of Models

This study constructed three separate models to explore the impact of convolutional layers. Model-A reduced the number of convolutional layers, Model-B and Model-C increased them compared to the baseline model. Fig. 1 indicates that Model-A which has reduced the convolutional layers can focus on broader features within the image. Conversely, increasing the number of convolutional layers led to a model that focused on finer, more localized image features.

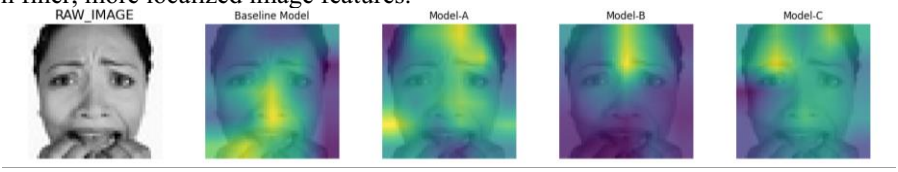

Fig. 1. Grad-CAM for Model-A, Model-B, and Model-C compared with baseline

Model-D, Model-E, and Model-F are used to explore the impact of pooling layers on model interpretability. From Model-D to Model-F, the number of pooling layers progressively decreases, with the pooling layers completely removed in Model-F. In terms of Fig. 2, it becomes evident that reducing the number of pooling layers in a model result in more discrete hotspots of attention.

Fig. 2. Grad-CAM for Model-D, Model-E, and Model-F compared with baselineDiscussion

In Model-A, Model-B, and Model-C, the convolutional layers have been adjusted by increase or decrease. Model-A is the only one that has decreased the convolutional and has a very different performance where the hotspots area is getting bigger. Model-B and Model-C have become highly focused on concentrated, with a severe lack of attention to non-hotspot areas. Too many convolutional layers can cause the model to focus on a single point or a few points, losing macro features and the perception of the entire field of view. In Model-D, Model-E, and Model-F, the pooling layers are being gradually removed. With the removal of the pooling layer, it is easy to observe that the hotspots in Grad-CAM become more dispersed, gradually shifting focus to more specific locations such as lips and eyes. This granularity enhanced detail recognition but possibly at the expense of losing broader contextual information, affecting its overall performance in varying scenarios compared to the baseline model. It should be noted that removing pooling layers significantly increases the model's parameters and computational demands, which is disadvantageous for training and application.

4 Conclusion

This study constructed various CNN models to examine the Grad-MAP of different models and explored how different layers impact model interpretability. The results demonstrate that adjusting the number of convolutional and pooling layers significantly affects the focus areas of the model. Specifically, convolutional layers influence the size of the focus areas, while pooling layers affect their dispersion. This finding provides new insights and practical guidance for the further development of facial expression recognition technology. Future research will focus on the suboptimal hotspot regions identified in this study, aiming to investigate the causes of these deviations.

References

1. Frank, M. G.: Facial expressions. In: International Encyclopedia of the Social & Behavioral Sciences, pp. 5230-5234. Elsevier (2001).
2. 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).
3. Liu, Y., Bao, Y.: Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. Measurement 220, 113418 (2023).
4. Liu, Y., Bao, Y.: Review on automated condition assessment of pipelines with machine learning. Advanced Engineering Informatics 53, 101687 (2022).

5. Zhao, F., Yu, F., Trull, T., Shang, Y.: A new method using LLMs for keypoints generation in qualitative data analysis. In 2023 IEEE Conference on Artificial Intelligence (CAI), pp. 333-334. IEEE (2023).
6. Qiu, Y., Wang, J.: A Machine Learning Approach to Credit Card Customer Segmentation for Economic Stability. In Proceedings of the 4th International Conference on Economic Management and Big Data Applications, ICEMBDA 2023, October 27-29, 2023, Tianjin, China, (2024).
7. Özay Ezerceli, M., Taner Eskil.: Convolutional Neural Network (CNN) Algorithm Based Facial Emotion Recognition (FER) System for FER-2013 Dataset (2022).
8. Chen, G., Zhang, D., Xian, Z., Luo, J., Liang, W., Chen, Y.: Facial Expressions Classification based on Broad Learning Network (2022).
9. Kaggle: FER-2013. Available at https://www.kaggle.com/datasets/msambare/fer2013 (2013).
10. 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).
11. Liu, Y., Yang, H., Wu, C.: Unveiling patterns: A study on semi-supervised classification of strip surface defects. IEEE Access 11, 119933-119946 (2023).
12. Wang, S., Zhang, Y.: Grad-CAM: Understanding AI Models. Computational Materials Continua 76(2), 1321-1324 (2023).