REVIEW OF FACIAL EXPRESSION RECOGNITION METHODS

Qixin Ye

University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

Facial expression recognition has gained significant attention in the field of computer vision and affective computing due to its potential applications in human-computer interaction, emotion analysis, and social robotics. In this paper, we introduce and compare three novel approaches for facial expression recognition: Attention-based Emotion Region CNN (AER-CNN), Multi-Scale Feature Fusion CNN (MSFF-CNN), and Novel FER Model. The AER-CNN model incorporates attention mechanisms for improved emotion recognition, while MSFF-CNN leverages multiple scales of features for better visual understanding. The Novel FER Model, on the other hand, incorporates a Dynamic Receptive Field (DRF) layer and a Multi-Scale Attention Module (MSAM) for adaptive and multi-scale analysis. We evaluate the performance of these models on the FER2013 dataset, with the Novel FER Model achieving the best validation accuracy of 50.10%. Despite efforts to mitigate overfitting through dropout layers and L2 regularization, some models still suffer from overfitting, possibly due to the low-quality 48x48 pixel images. The introduction of the DRF layer shows promise in improving the accuracy of facial expression recognition models.

Index Terms— Facial Expression Recognition, CNN, Attention Based, Multi-Scale Attention, Dynamic Receptive Field

1. INTRODUCTION

Facial expression recognition has gained significant attention in the field of computer vision and affective computing due to its potential applications in fields such as human-computer interaction, emotion analysis, and social robotics. Recognizing and understanding facial expressions from images or videos poses several challenges, such as capturing subtle facial cues, handling variations in expression intensity, and dealing with diverse facial appearances and scales.One popular approach is to use convolutional neural networks (CNNs) to automatically extract relevant features from visual data. However, traditional CNNs often lack the ability to capture the fine-grained emotion-related information that may be distributed across different regions of an image or video. To address this limitation, three novel approach were introduced, which were Attention-based Emotion Region CNN (AER-CNN), Multi-Scale Feature Fusion CNN (MSFF-CNN) and Novel FER Model.

2. MODEL ARCHITECTURES

Attention-based Emotion Region CNN (AER-CNN)[1] incorporates attention mechanisms, which allow the model to dynamically focus on different regions of an image or video that are more relevant to emotions. This attention-based approach enables AER-CNN to capture emotional cues from specific regions of visual stimuli, resulting in improved emotion recognition performance. In this paper, we will provide an overview of AER-CNN and its key components, highlighting its potential applications in various fields such as affective computing, human-computer interaction, and multimedia analysis. We will also review the current state-of-the-art approaches in emotion recognition, and discuss the advantages and limitations of AER-CNN. Finally, we will provide insights on future research directions and potential advancements in the field of attention-based emotion region CNNs.

Multi-Scale Feature Fusion CNN(MSFF-CNN)[2] leverages multiple scales of features to capture both global context and local details, leading to improved performance in various computer vision tasks. MSFF-CNN incorporates feature fusion techniques that allow the model to integrate information from multiple scales, enabling it to capture both fine-grained and coarse-grained features in visual stimuli. In this paper, we provide an overview of MSFF-CNN and its key components, highlighting its potential applications in tasks such as image classification, object detection, and semantic segmentation. We review the current state-of-the-art approaches in feature fusion and multi-scale modeling, and discuss the advantages and limitations of MSFF-CNN. Furthermore, we provide insights on future research directions and potential advancements in the field of multi-scale feature fusion CNNs, which have shown promising results in addressing the challenges of capturing information at multiple scales for improved visual understanding.

Novel FER Model incorporates a Dynamic Receptive Field (DRF)[3] layer and a Multi-Scale Attention Module (MSAM). The DRF layer is designed to adaptively adjust the receptive field of a neural network based on the input content, allowing the network to capture contextual information at

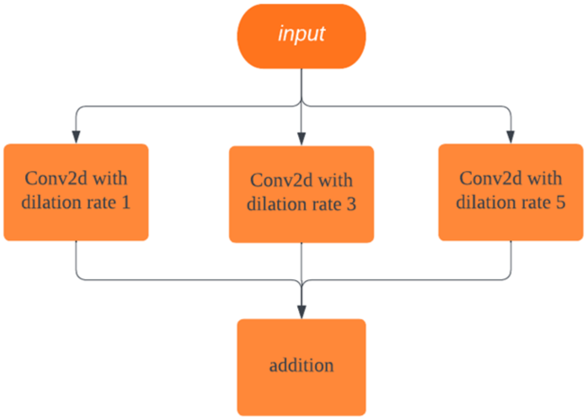
input
Conv2d with dilation rate 1
Conv2d with dilation rate 3
Conv2d with dilation rate 5
addition

Fig. 1. The structure of DFR model

different scales. Unlike traditional fixed receptive fields, the DRF layer dynamically adjusts the size of the receptive field based on the content of the input, allowing the network to focus on both local details and global context. This adaptability makes the DRF layer particularly useful in capturing contextual information from objects or scenes with varying scales, which is common in real-world images or videos.

The Multi-Scale Attention Module(MSAM)[4], on the other hand, is designed to incorporate multi-scale attention mechanisms into a neural network. Attention mechanisms have been widely used in deep learning to selectively focus on informative regions of an input. The MSAM takes this concept to the next level by incorporating attention mechanisms at multiple scales, allowing the network to selectively attend to different levels of features. This enables the MSAM to capture both fine-grained details and coarse-grained context information, leading to improved performance in tasks that require multi-scale analysis.

In this paper, we compare these three advanced methods, namely AER-CNN, MSFF-CNN,and Novel FER Model, specifically in the context of facial expression recognition. We highlight the key components and working principles of these methods, and discuss their potential applications in facial expression recognition tasks. We review the current state-of-the-art approaches in facial expression recognition, and discuss the advantages and limitations of these methods. Furthermore, we provide insights on potential future research directions and advancements in the field of facial expression recognition using AER-CNN, MSFF-CNN, DRF, and MSAM, which have shown promising results in addressing the challenges of facial expression analysis in complex and diverse real-world scenarios.

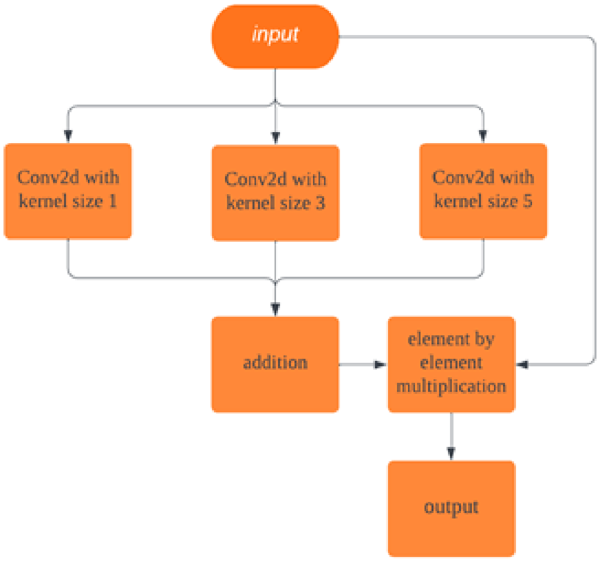
input
Conv2d with kernel size 1
Conv2d with kernel size 3
Conv2d with kernel size 5
addition
element by element multiplication
output

Fig. 2. The MSAM model

3. EXPERIMENT PROCESS AND RESULTS

3.1. Data Preprocessing

In this work, we use the FER2013 dataset for Facial Expression Recognition (FER) problem. The dataset contains 32,298 grayscale images of size 48×48 pixels, with 7 different emotion labels: angry, disgust, fear, happy, sad, surprise, and neutral.

Before feeding the images into a deep learning model, we perform several preprocessing steps to enhance the quality and consistency of the data. First, we count the number of images for each emotion in the train and test sets to ensure that the data is balanced. This is important because an imbalanced dataset can lead to biased model performance.

Next, we preprocess each image using the following steps:

- **Normalization:** We normalize the pixel values to a range of 0-255 using the following formula:

$$x' = \frac{(x - \min(x)) \times 255}{(\max(x) - \min(x))} \tag{1}$$

  where x is the original pixel value, x' is the normalized pixel value, $\min(x)$ and $\max(x)$ are the minimum and maximum pixel values in the image, respectively. This step ensures that all images have the same range of pixel values, which can help the model learn more effectively.

- **Histogram Equalization:** We apply histogram equalization to each image to improve contrast and enhance the visibility of important features[5]. Histogram equalization is a technique that redistributes the pixel values in an image to achieve a more uniform distribution. It is defined by the following formula:

$$g(i,j) = \frac{255}{N \times M} \sum_{k=0}^{i} \sum_{l=0}^{j} h(k,l) \tag{2}$$

where $g(i,j)$ is the equalized pixel value at position (i,j) in the image, $h(k,l)$ is the histogram of the original image up to position (k,l), and N and M are the dimensions of the image. This step can help the model better distinguish between different facial expressions by highlighting key features such as facial landmarks and facial expressions.

- **Noise Removal:** Finally, we remove noisy images with low variance using the following criteria:

$$\operatorname{var}(I) \geq T \tag{3}$$

where $\operatorname{var}(I)$ is the variance of the pixel values in the image I, and T is a threshold value. We set the threshold to 10 based on empirical results, as images with variance below this threshold were found to contain little useful information for the model to learn from.

3.2. Attention-based Emotion Region CNN (AER-CNN)

Our first model, AER-CNN, employs attention mechanisms to emphasize relevant features. We train the model using the Adam optimizer with a learning rate of 0.001 for 25 epochs. The training accuracy reaches 51.23%, while the validation accuracy reaches 45.70%.

3.3. Multi-Scale Feature Fusion CNN (MSFF-CNN)

In the second model, Multi-Scale Feature Fusion CNN (MSFF-CNN), we initially face an overfitting issue, which is evident from the high training accuracy of 71.46% and a much lower validation accuracy of 49.98% after 5 epochs. To mitigate this problem, we introduce dropout layers and L2 regularization. (In Fig. 3.)

3.3.1. Dropout Layers

Dropout is a regularization technique that helps prevent overfitting in deep learning models by randomly setting a fraction of input units to 0 during training. This prevents the model from relying too heavily on any single input feature, thereby encouraging it to learn more robust and generalized representations[6].

In the MSFF-CNN model, we add dropout layers with a 30% dropout rate after every batch normalization layer in each of the three branches (x1, x2, x3). This addition slightly improves the overfitting issue, yielding a training accuracy of 65.98% and a validation accuracy of 48.92% after 5 epochs. (In Fig. 4.)

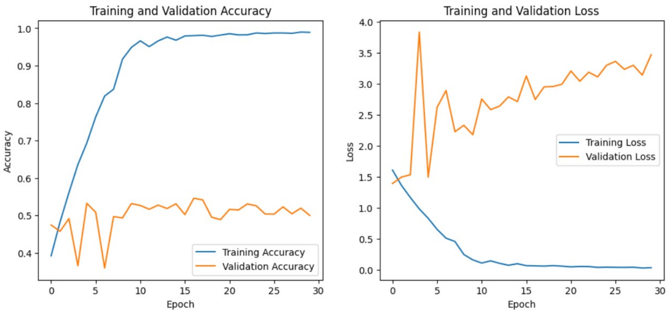

Fig. 3. MSFF-CNN

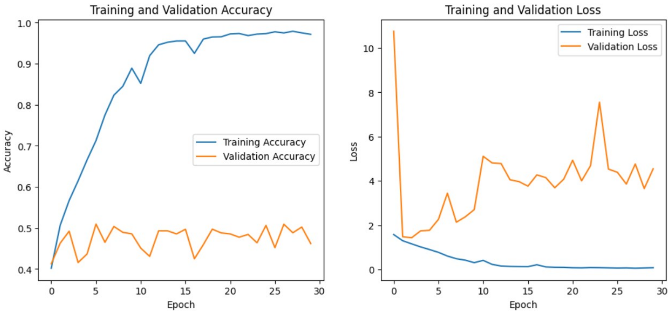

Fig. 4. MSFF-CNN with dropout layers

3.3.2. L2 Regularization

L2 regularization, also known as weight decay, is another technique used to reduce overfitting in neural networks. It works by adding a regularization term to the model's loss function, which penalizes large weights[7]. This encourages the model to learn simpler and more generalizable patterns in the data.

We further apply L2 regularization to the MSFF-CNN model in an attempt to improve its performance. After incorporating L2 regularization, we observe a moderate improvement in overfitting. The training accuracy reaches 62.39%, and the validation accuracy reaches 44.79% after 5 epochs. (In Fig. 5.)

Despite the improvements achieved through dropout layers and L2 regularization, the MSFF-CNN model still suffers from overfitting. As a result, we ultimately discard this model and opt for the Novel FER Model, which demonstrates better performance with a training accuracy of 56.88% and a validation accuracy of 50.10%.

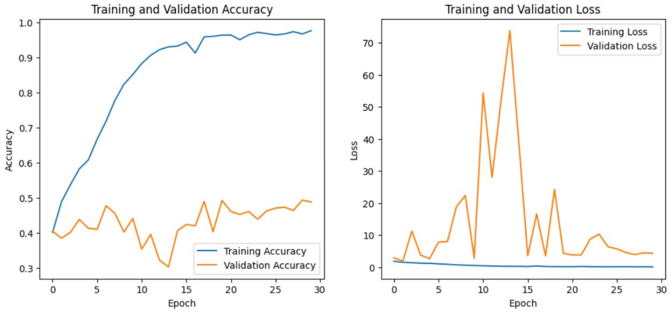

Fig. 5. MSFF-CNN with L2 regularization and dropout layers

3.4. Novel FER Model

Our third model incorporates a Dynamic Receptive Field (DRF) layer and a Multi-Scale Attention Module (MSAM). We train this model using the Adam optimizer with a learning rate of 0.002 for 100 epochs. The training accuracy reaches 56.88%, and the validation accuracy reaches 50.10%. (In Fig. 6.)

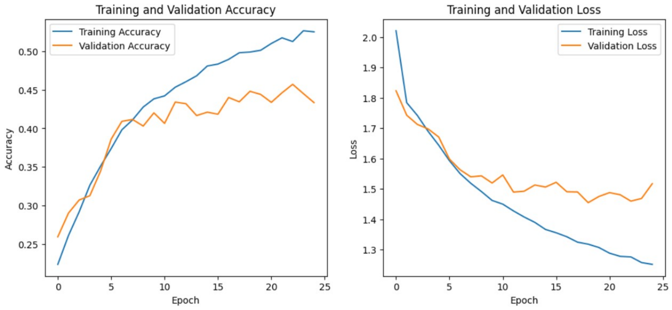

Fig. 6. Novel FER Model

3.5. Results

The third model (Novel FER Model) demonstrates the best performance among the three models, with a training accuracy of 56.88% and a validation accuracy of 50.10%. The second model (MSFF-CNN) highlights the benefits of using dropout layers and L2 regularization to mitigate overfitting, although the model still suffers from overfitting and is ultimately discarded.

4. DISCUSSION

Despite achieving an accuracy of up to 50%, overfitting remains a challenge for some of our models. The small size (48 by 48 pixels) of our images may have contributed to this issue, as well as the high number of parameters in the neural network. Although we attempted to address this through the use of dropout and L2 regularization, the results were not as significant as we had hoped. Nonetheless, our results indicate that the reception field layer is a promising technique for enhancing accuracy compared to MSFF and AER-CNN.

In a similar study, Zhao et al. employed a comparable CNN architecture with multi-scale and local attention features for FER2013[8]. However, their CNN had more parameters and a validation set accuracy of approximately 59.4%. A drawback of our model in comparison to theirs is that it cannot distinguish between facial expressions and non-facial expression images. Furthermore, while Zhao et al. trained their model on multiple datasets, we focused solely on the FER2013 dataset. Our approach introduced the concept of the reception field layer, which enables the model to extract global information directly from the image, an approach that was not implemented in Zhao et al.'s model.

5. CONCLUSION

This paper proposed three different CNN architectures, namely AER, MSFF, and Novel FER-CNN, to address the challenge of facial expression recognition. Our results showed that the AER-CNN approach can achieve an accuracy of up to 45.70%, MSFF-CNN has approximately 49.98%, and the Novel-FER method achieves the highest accuracy of 50.10%. Overall, our focus on attention and global learning was effective in identifying crucial information in the image, enabling the output of facial expressions. Further research could explore the use of larger images and a larger dataset to mitigate the overfitting issue and improve accuracy.

6. REFERENCES

[1] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1771–1775.

[2] M. Swathy, S. Logesh Kumar, T. Arunkumar, G. Lavanya, and R. Saranya, "Identification of bone fracture lesions in digital x-ray images using msff and msfd method," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 2022, pp. 208–213.

[3] Amin Fakhartousi, Sofia Meacham, and Keith Phalp, "Autonomic dominant resource fairness (a-drf) in cloud computing," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022, pp. 1626–1631.

[4] Dongdong Cui, Shouyi Yin, Jiangyuan Gu, Leibo Liu, and Shaojun Wei, "Msam: A multi-layer bi-lstm based

speech to vector model with residual attention mechanism," in *2019 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, 2019, pp. 1–3.

[5] H.D. Cheng and X.J. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digital Signal Processing*, vol. 14, no. 2, pp. 158–170, 2004.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[7] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, pp. 022022, feb 2019.

[8] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021.