Scalable Facial Landmark-based Emotion Recognition with Multi-Branch Attention

Savina Jassica Colaco and Dong Seog Han*
School of Electronic and Electrical Engineering
Kyungpook National University, Daegu, Republic of Korea
savinacolaco@knu.ac.kr,* dshan@knu.ac.kr

***Abstract*—Automatic emotion recognition is crucial for a wide range of applications, including medical imaging, virtual reality, and security. The best performance in face emotion identification has yet to be reached despite new deep learning algorithms generating great results. This paper proposes a deep learning model for facial emotion recognition that combines mini Xception modules and attention mechanisms with facial landmarks. The model is further improved with multi-branch attention for emotion classification The model is trained using the FER2013 benchmark dataset and achieves 53.58% classification accuracy.**

***Index Terms*—Convolutional neural network (CNN), emotion classification, facial landmarks**

I. Introduction

Human emotions are the standard mode of human interaction universally. It has been studied for years to develop robust automatic emotion recognition systems. Emotion recognition systems are used in various applications such as self-driving monitoring, security systems, health care systems, entertainment systems like gaming, etc. Emotion recognition has been developed with consideration of different modalities such as audio signals, biosignals and video streams [1]. To identify any facial expression, facial landmarks help to describe various human expressions. The facial landmarks are essential points on the face, such as the eyes, nose, mouth, eyebrows and jawline. Some primary emotions, such as happy, sad, neutral, surprise, etc, are depicted by human faces with slight changes in muscle movement. Hence, these changes can be captured by facial landmarks to detect emotions in automatic systems. Several methods fuse landmark information with image features to improve detection, but it has yet to be studied in recent years. Recent emotion recognition has been developed using deep neural networks for the identification of complex human expressions which are inaccurately identified by traditional methods.

In this paper, we propose a deep neural network for classifying human emotions. The feature learning module, such as Xception [2], learns both image and landmark features for better classification of emotions. Since there exists the dependence of some facial points on other emotions, the multi-branch attention method is utilized to focus on essential points of the face. These attention-based features help to improve the emotion classification performance. The model is lightweight due to depthwise convolution layers.

The rest of the paper is organized as follows: Section II provides the related works on image classification. The proposed models are explained in detail in Section II. Section III provides the results of the experiments and their discussion. Finally, we conclude in Section IV.

II. Proposed Model

Scaling the standard convolution layer could involve scaling the width or depth of the network. Scaling either of them could improve the accuracy or encounter a vanishing gradient problem [3]. Compound model scaling allows scaling of width, depth and image resolution of the network uniformly. We propose a scalable model to avoid overfitting of the model, which is encountered with scaling only one of the parameters. The proposed model contains image and landmark feature learning subnetworks. These subnetworks are mini Xception modules that use depthwise separable convolutions with skip connections. Batch normalization BN and the ReLU activation function are applied after each convolution and depthwise separable convolution layer $SepConv$. The mini Xception modules contain depthwise convolutions, which are not followed by pointwise convolutions. Each subnetwork is given different inputs, i.e. image and landmark input. The output from both feature learning subnetworks is further fed to the attention module. The attention mechanism contains spatial attention SA and channel attention CA in end-to-end design. In spatial attention, the mean and standard deviation of input are calculated, fed to the convolution layer to generate attention weights and multiplied with input. In channel attention, the global average pool of the input is calculated. This is fed to the convolution layer to get attention weights and multiplied with input. All the outputs from attention are concatenated and passed to multi-branch attention. Multi-branch attention consists of convolution layers with different kernel sizes to extract multiscale features under different receptive fields. The different kernel sizes are 1×1, 3×3 and 5×5. Each is followed by spatial and channel attention to learn important presentations at higher levels of the network. The residual structure with 1×1 convolution is utilized to enhance multiscale depth. The width w and depth d of the whole network are scaled uniformly. The output from multi-branch attention is passed to a fully connected FC layer to classify the

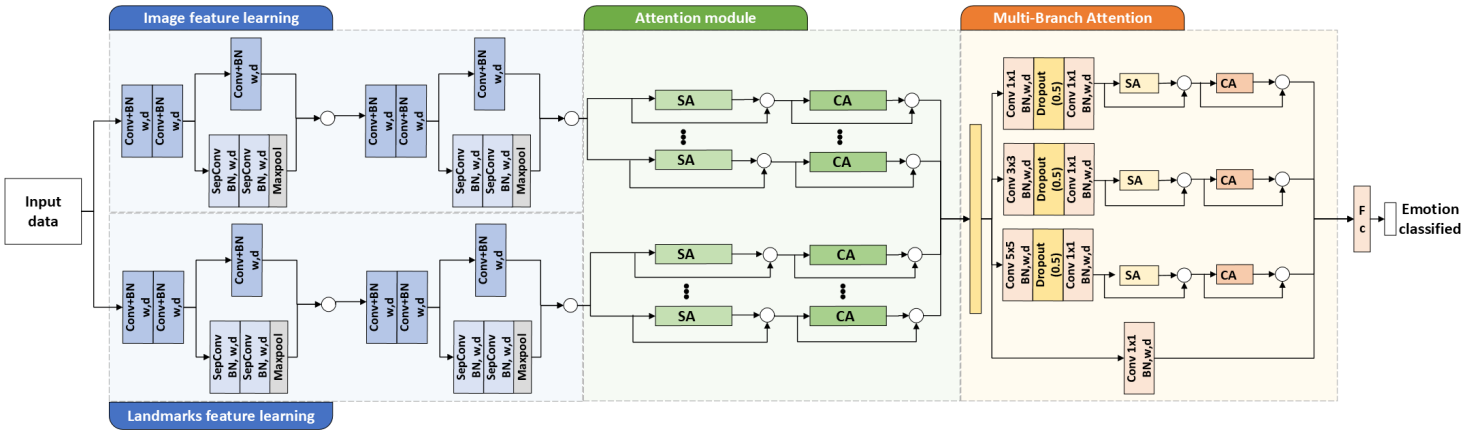

Fig. 1. Proposed model for emotion recognition.

emotions. Fig. 1 provides the detailed structure of the proposed model.

III. Results and Discussion

A. Implementation Details

The proposed model was trained using the FER2013 [4] dataset, which contains 28,709 training images and 7,178 test images divided into seven emotion classes. The test data is divided equally into two categories: test data and unseen data. angry, disgust, fear, happy, sad, surprise and neutral are the emotion classes. The images have been resized to 48×48 pixels in size. The model is trained using a batch size of 32 and epochs of 100. The model is optimized with the Adam optimization technique with a learning rate of 10^{-3}. The learning rate is reduced every 10 epochs, and the best values are saved based on loss values. The categorical cross-entropy [5] loss is used as the loss function.

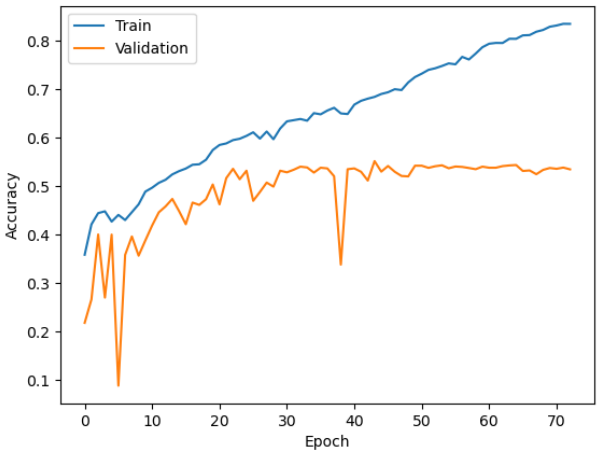

Fig. 2. Accuracy of the proposed model.

B. Discussion

The performance of the proposed model has been evaluated with accuracy and a confusion matrix. It helps to understand the significance of landmarks with a multi-branch attention module. The landmarks are acquired using the dlib library, which returns 68-point landmarks for each image in the dataset. They are utilized to capture essential features for emotion recognition. From the accuracy plot in Fig. 2, the validation accuracy of the model achieved is around 53.58%. The model achieves a similar test accuracy of around 53.10%. The model [6] trained with images and landmarks and width and depth values is 0.5 and has a lower accuracy of 50% compared to the proposed model.

The utilization of spatial and channel attention helped the model to focus on essential features that are needed for emotion classification. As the compound model scaling method scales the width and depth uniformly, it prevents the model from overfitting the model. The labels of the confusion matrix in Fig. 3 are as follows: angry: 0, disgust: 1, fear: 2, happy: 3, sad: 4, surprise: 5, and neutral: 6. The depth and width values considered are 0.5 to reduce the model size. According to the confusion matrix, the proposed model trained with images and landmarks recognizes labels with significant facial variations, such as happy and surprise. The happy labels have the highest data; hence, it has better classification. It slightly misclassifies with angry and disgust labels since they exhibit similar facial movements. The model trained with both images and landmarks has improved accuracy with both train data and unseen data. From the confusion matrix, we notice that the emotion labels with the most minor facial variations, such as sad and neutral, are often confused due to similar facial variations. Hence, landmarks help to detect the emotion adequately. Also, the multi-branch attention captures features at different receptive fields with essential representations, which are helpful for understanding emotion representations. Though the gap between validation accuracy and test (unseen data) accuracy is small, the model still needs to improve its accuracy.

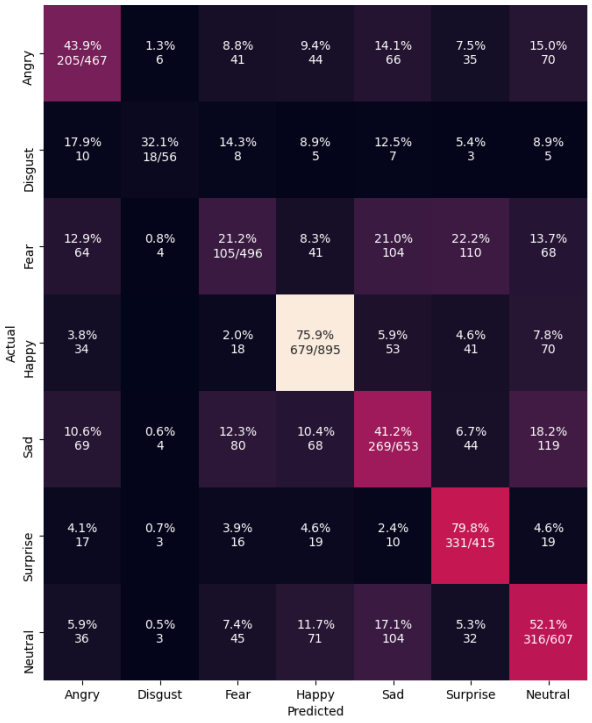

Fig. 3. Confusion matrix of proposed model.

IV. Conclusion

This paper proposed a model consisting of subnetwork feature learning of both images and landmarks. These subnetworks are mini scalable Xception modules with attention mechanisms. The model is improved with multi-branch attention to capture multi-scale important features The model utilizes landmarks and images to detect emotions. The model's classification accuracy on unseen data is improved, but it requires further enhancement for increased generalizability. In the future, it will be feasible to concentrate on particular areas of the face to identify simple and complex emotions because not all facial features are required for emotion recognition.

Acknowledgment

This work was supported by the Technology Development Program (RS-2023-00222555), funded by the Ministry of SMEs and Startups(MSS, Korea).

References

[1] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," IEEE transactions on cybernetics, vol. 49, pp. 839-847, 2018.

[2] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[3] S. J. Colaco and D. S. Han, "Deep Learning-Based Facial Landmarks Localization Using Compound Scaling," IEEE Access, vol. 10, pp. 7653-7663, 2022.

[4] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., "Challenges in representation learning: A report on three machine learning contests," in Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013.Proceedings, Part III 20, 2013, pp. 117-124.

[5] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," in IEEE Access, vol. 8, pp. 4806-4813, 2020, doi: 10.1109/ACCESS.2019.2962617.

[6] S. J. Colaco and D. S. Han, "Analysis of Scalable Attention-based Emotion Recognition with Facial Landmarks" Proceedings of the Korean Society of Communication Studies, (2023): 1298-1299.