INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Facial Expression Recognition Using Convolutional Neural Network with Attention Module

Habib Bahari Khoirullah [a], Novanto Yudistira [a,*], Fitra Abdurrachman Bachtiar [a]

[a] *Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, East Java, 65145, Indonesia*

*Corresponding author: *yudistira@ub.ac.id*

***Abstract*—Human Activity Recognition (HAR) is an introduction to human activities that refer to the movements performed by an individual on specific body parts. One branch of HAR is human emotion. Facial emotion is vital in human communication to help convey emotional states and intentions. Facial Expression Recognition (FER) is crucial to understanding how humans communicate. Misinterpreting Facial Expressions can lead to misunderstanding and difficulty reaching a common ground. Deep Learning can help in recognizing these facial expressions. To improve the probation of Facial Expressions Recognition, we propose ResNet attached with an Attention module to push the performance forward. This approach performs better than the standalone ResNet because the localization and sampling grid allows the model to learn how to perform spatial transformations on the input image. Consequently, it improves the model's geometric invariance and picks up the features of the expressions from the human face, resulting in better classification results. This study proves the proposed method with attention is better than without, with a test accuracy of 0.7789 on the FER dataset and 0.8327 on the FER+ dataset. It concludes that the Attention module is essential in recognizing Facial Expressions using a Convolutional Neural Network (CNN). Advice for further research first, add more datasets besides FER and FER+, and second, add a Scheduler to decrease the learning rate during the training data.**

***Keywords*—Facial expression recognition; attention; CNN.**

Manuscript received 15 Jun. 2022; revised 5 Aug. 2022; accepted 29 Sep. 2022. Date of publication 31 Dec. 2022.

I. INTRODUCTION

Human Activity Recognition (HAR) is an introduction to human activities that refer to the movements performed by an individual on certain body parts. HAR has become a widely discussed scientific topic in the Computer Vision community because it is involved in many Human-Computer Interaction (HCI) application developments [1], [2]. One branch of HAR is human emotion. Facial emotion is vital in human communication to help convey emotional states and intentions [3]. Communication is exchanging information between individuals or groups with the meaning or purpose to be conveyed. The message or information conveyed can be in the form of verbal communication or non-verbal communication. Among these non-verbal components, facial emotions are essential in communication [4].

Facial expressions display personal emotions and indicate individual intentions in social situations; therefore, they are crucial for social communication. However, many previous studies have explored the processing of isolated facial expressions without communication with other people, whereas humans rarely interact directly with faces without context [5], [6].

Facial Expression Recognition (FER) is a technique to understand human emotions from the expressions they display as a reaction to something that occurs in the environment. This technique can be used in photos, videos, and real-time [7]. FER can be widely applied today, such as in understanding human expressions in online meetings using video. At online meetings, there are difficulties in the interaction between participants, which can lead to misunderstandings and difficulty in digesting a conversation that causes misunderstandings in meeting participants[8]. There is a model in the neural network, namely attention, which imitates cognitive attention. The effect increases the crucial parts of the input data and eliminates the rest. Which part of the data is more important than the others depends on the context and is studied through training data with gradient descent [9].

Many studies have been carried out for FER. However, for data from Affect Net, the highest accuracy obtained is only 59.5% using the ResNet18 model, a deep learning model that functions for classification [10]. Therefore, we need a robust architecture to obtain the best classification for FER. In this study, two CNN architectures were evaluated to achieve the best results: ResNet 50, and VGGFace. ResNet 50 was used because it has better precision than ResNet18, and VGGFace was used due to the accuracy reached 88% [11].

II. Materials and Method

A. The Dataset

The dataset used in this study is images of Facial Expressions FER and FER+ by Microsoft [12], consisting of various human facial expressions. The FER dataset has a total of 35.887 images split into eight labels, which can be seen in Fig. 1, which consist of Happy, Angry, Neutral, Surprise, Fear, Disgust, and Sad. The FER+ dataset consists of ten labels (Happiness, Sadness, Anger, Contempt, Neutral, Surprise, Fear, Disgust, Unknown, and Nf). Each represents a human expression taken through a controlled environment, providing us with a black-and-white image. Image data of a 48x48 pixel are provided as a grayscale image of the face. Faces are automatically registered, placed almost in the center, and each frame occupies about the same space. The task is to classify each face based on the emotions shown in the facial expressions [13].

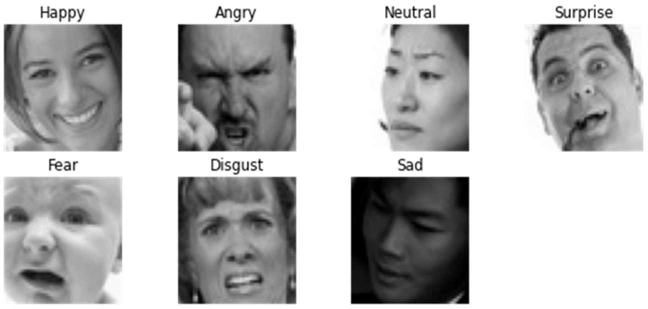

Fig. 1 FER dataset samples.

The difference between FER and FER+ is that the FER+ is the new relabelled dataset by ten crowd-sourced taggerregis, as can be seen in Fig. 2, which proved better quality ground truth for still image emotion than the original FER labels[14]. Ten taggers for each image enable researchers to estimate an emotion probability distribution per face, and this allows for constructing algorithms that produce statistical distributions or multi-label outputs instead of the conventional single-label output. The split distribution in this research for the dataset was split into 80% training set and 20% for the testing set. This approach of dataset split assured that these models performed well on the unknown data instead of the usual method of only a 10% testing set.

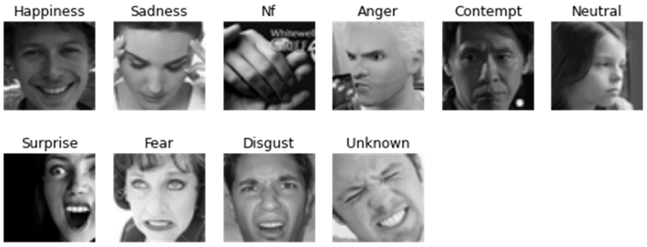

Fig. 2 FER+ dataset samples.

B. Proposed Method

Convolutional Neural Network (CNN) is one of the machine learning methods developed by MultiLayer Perceptron (MLP), which is designed to process two-dimensional data [15]. CNN consists of two main stages, namely feature learning and classification. Feature learning consists of a convolution layer, pooling layer, and ReLu, while classification consists of flattening, fully connected, and softmax [16].

Residual Network (ResNet)[17] is an architecture of CNN invented in 2015 and has won the ILSVRC 2015 with the top error of 3.57. This is because ResNet uses a residual network, reducing complexity and solving the degradation while maintaining good performance. It is one of the architectures of deep learning.

ResNet was chosen because it has a residual connection mechanism, a form of connection in an artificial neural network created by adding a shortcut between two points. Adding a shortcut to the ResNet architecture allows the optimization method to update the weights on the gradient in the previous layer so that the initial layer can be updated with better weights[18]. Fig. 3 concisely shows how the residual layer works.

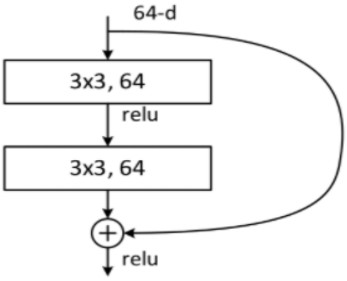

Fig. 3 ResNet Residual Block

ResNet has shown that robust performance still needs improvements. Spatial Transformer [19] also has an attention mechanism that explicitly allows spatial data manipulation in a network. This different module can be embedded in an existing convolution architecture, allowing neural networks to actively modify feature maps spatially, conditionally on the feature map itself, without training supervision or additional modifications to the optimization process. We show that using spatial transformers produces invariance for translation, scaling, rotation, and warping, yielding the most advanced performance across several benchmarks and transformation classes. The architecture of deep learning in Fig. 4 shows the attention mechanism, which considers the localization process before processing it through a grid generator.

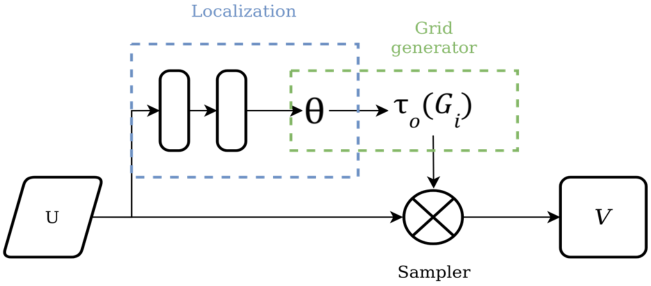

Fig. 4 Attention Mechanism in Spatial Transformer Network.

The attention mechanism in spatial transformer network consist of a localization layer, which takes the localization network takes the input feature map U ∈ $\mathbb{R}^{\boldsymbol{HWC}}$with width W, height H and C channels and outputs $\boldsymbol{\theta}$. The localization network function can be in any shape, such as a fully connected or convolutional network. However, it must have a final regression layer to provide the transformation parameters. In this research, we use 2 Convolutional layers of 64 by 64 instead of the fully connected layers for the localization network since Convolutional layers perform better than the usual fully connected layers. There is also the Grid generator in that each output pixel is computed by using a sampling kernel centered at a particular place in the input feature map to execute a warping of the input feature map such that the $\boldsymbol{\tau_o(G_i)}$ is formulated as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \tau_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (1)$$

Where (x_i^t, y_i^t) is the target coordinates of the normal grid of the output feature map, (x_i^s, y_i^s) is the source coordinates of the input feature map that defines the sample points, and A_θ is the affine transformation θ matrix which might take various transformations. Normalized height and width coordinates so that $-1 \le x_i^t, y_i^t \le 1$ are within the spatial boundaries of the output and $-1 \le x_i^s, y_i^s \le 1$ are within the spatial boundaries of the input[20]. We can take this mechanism and combine it with ResNet, as seen in Fig. 5,

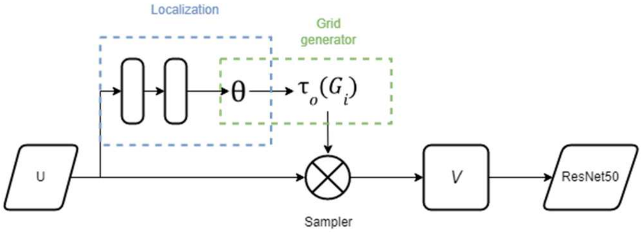

Fig. 5 ResNet with Attention Mechanism.

While the normal attention in the Spatial Transformer network has the original untouched image in the sampler, this research added a ResNet model after the sampler mechanism. Although our model was much deeper and had more parameters since the sampler makes our dataset more varied from the affine transformation, ResNet might be able to pick up the different features of the face better.

III. Result and Discussion

A. Models Comparison

In this study, various models can be used to classify facial expression recognition. One way to obtain a model with the best results is to compare all existing models. In this study, one of the models considered in this experiment is ResNet50, which consists of 5 stages with a convolution and Identity block. Each convolution block has three convolution layers, and each identity block has three convolution layers. The ResNet-50 has over 23 million trainable parameters. On the other hand, VGGFace has five stages of the convolutional block, each of which has three convolutional layers, which is not much different from ResNet50. However, VGGFace has 145 million trainable parameters, around seven times more than ResNet50. Vanilla CNN as our baseline uses three stages of convolutional block. Each block has three convolution layers but only 2 million parameters, much less than the other two models we test. This research defines baseline hyperparameters, such as a learning rate of 0.001 and an Adam optimizer. From the results of training and testing on several models in Table I, it was found that ResNet50 obtained the highest accuracy for training and testing data, with VGGFace as the second-highest model and vanilla CNN being the lowest. Therefore, because the ResNet50 model produces higher accuracy than other models (VGGFace, and Vanilla CNN), the continuation of this research used the ResNet50 model.

TABLE I
Models comparison in accuracy

Model	Accuracy			
	Training		Testing	
	FER	FER+	FER	FER+
ResNet50	**0.864833**	**0.916081**	**0.771539**	**0.823987**
VGGFace	0.860744	0.913236	0.760199	0.812477
Vanilla CNN	0.819416	0.873216	0.755404	0.809351

B. Hyperparameters Tuning

In this study, various hyperparameters can be used to train facial expression recognition. One is optimizers, which are used to update the weights during training. We compare five optimizers, namely SGD, Adam, NAdam, RMSProp, and Radam, which can be seen in Table II.

From the results of training and testing on several models in Table II, it was found that the highest accuracy for FER data training was RAdam, with an accuracy of 0.883573, and for FER+ was NAdam, with an accuracy of 0.937226. However, the FER and FER+ test data results show that the Adam optimizer is the best, with 0.771539 on the FER test data and 0.823987 on the FER+ data.

TABLE II
Optimizers comparison in accuracy

Optimizer	Accuracy			
	Training		Testing	
	FER	FER+	FER	FER+
ResNet50 + SGD	0.390426	0.441912	0.485793	0.538031
ResNet50 + Adam	0.860744	0.913236	**0.771539**	**0.823987**
ResNet50 + NAdam	0.883262	**0.937226**	0.767787	0.822016
ResNet50 + RMSprop	0.871164	0.924262	0.741632	0.792464

It can also be seen in Fig. 6 to Fig. 9 that the Adam optimizer produces the optimization process converges at the tenth epoch. In contrast, other optimizers require more than a tenth of epochs to reach the converged graphs. Therefore, because the Adam optimizer yields the best results, the continuity of this research used the Adam optimizer.

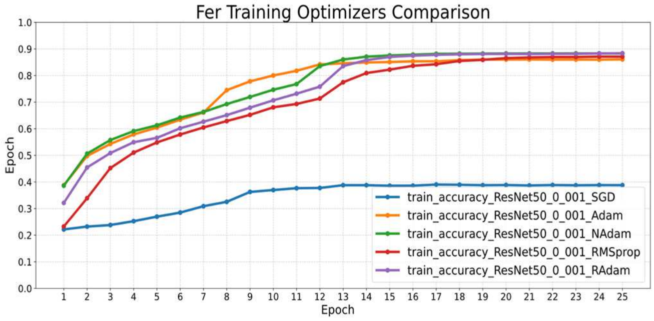

Fig. 6 Optimizers Comparison on Training Data of FER

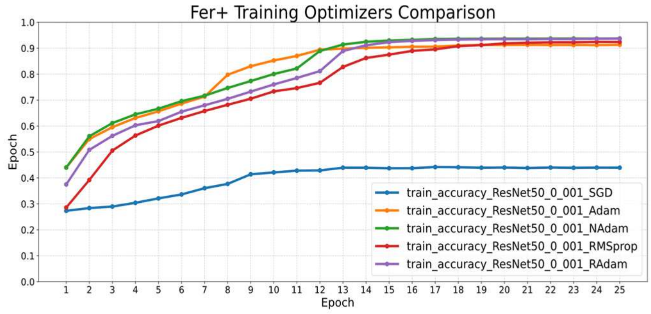

Fig. 7 Optimizers Comparison on Training Data of FER+

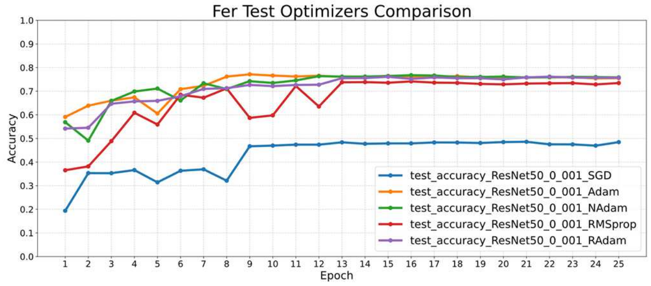

Fig. 8 Optimizers Comparison on Testing Data of FER

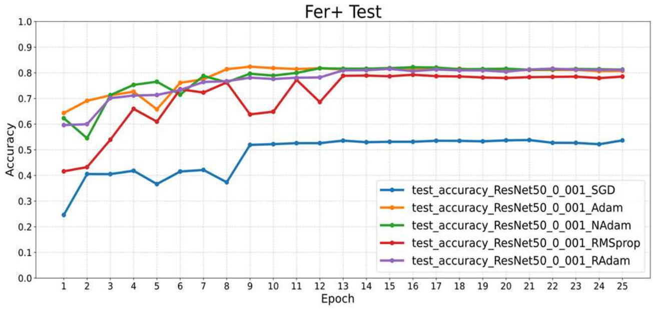

Fig. 9 Optimizers Comparison on Testing Data of FER+

C. Learning Rate Comparisons

In this study, various hyperparameters can be used to learn facial expression recognition. One of them is the learning rate. Using the Adam optimizer to optimize the ResNet50 model, we compare six learning rates, namely 0003, 0.0003, 0.001, 0.0001, 0.001, and 0.003, which can be seen in Table III. The training and testing results on several models in Table III show that a learning rate of 0.003 is the best result in training, which is 0.885839 on FER data, and 0.9376684 on FER+ data. However, accuracy on test data shows that the learning rate of 0.0001 is the best learning rate on FER and FER+ test data, with accuracy results of 0.771539 and 0.823987, respectively.

TABLE III
LEARNING RATE COMPARISON IN ACCURACY

Optimizer	**Accuracy**			
	Training		**Testing**	
	FER	**FER+**	**FER**	**FER+**
ResNet50 + Adam + 0.03	0.249039	0.301279	0.363293	0.416162
ResNet50 + Adam + 0.0003	**0.885839**	**0.937684**	0.739576	0.791338
ResNet50 + Adam + 0.01	0.803197	0.855602	0.732255	0.787116
ResNet50 + Adam + 0.0001	0.883957	0.936064	0.677038	0.730544
ResNet50 + Adam + 0.001	0.860744	0.913236	**0.771539**	**0.823987**

Fig. 10 to Fig. 13 show that a learning rate of 0.0001 can achieve maximum accuracy only at the tenth epoch, while for other learning rates, it requires more than the twelfth epoch to get the highest accuracy value. Therefore, from the results of the learning rate test, a learning rate of 0.0001 was used to continue this research.

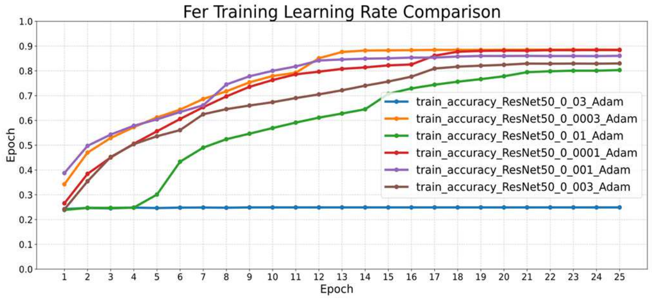

Fig. 10 Learning Rates Comparison on Training Data of FER

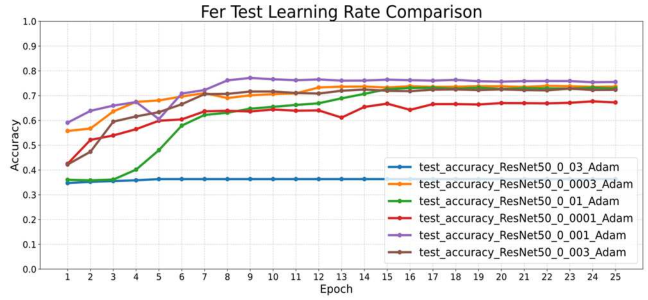

Fig. 11 Learning Rates Comparison on Testing Data of FER

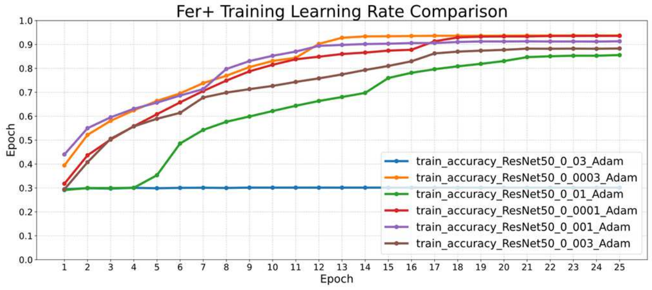

Fig. 12 Learning Rates Comparison Training Data of FER+

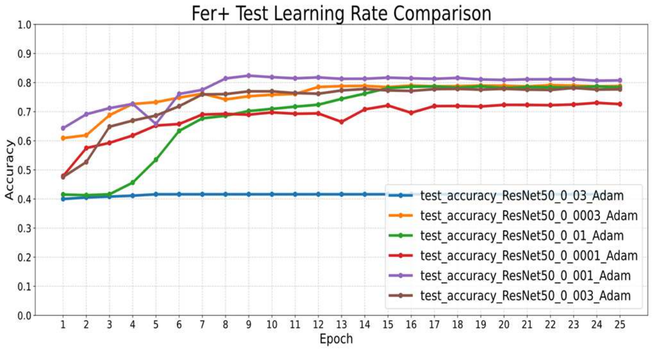

Fig. 13 Learning Rates Comparison on Testing Data of FER+

D. Attention Comparisons

In this study, we will compare ResNet50 without and with the attention module with the best optimizer and learning rate that has been obtained from previous experiments. The results of using the attention mechanism can be seen in Table IV. The table concludes that the attention module contributes to increasing the performance, with accuracy on training and test data on FER and FER+ are better than that do not use the attention mechanism.

TABLE IV
ATTENTIONS COMPARISON IN ACCURACY

Optimizer	**Accuracy**			
	Training		**Testing**	
	FER	**FER+**	**FER**	**FER+**
ResNet50 + Adam + 0.001	0.860744	0.913236	0.771539	0.823987
ResNet50 + Adam + 0.001 + Attention	0.895158	0.947883	0.778905	0.832741
ResNet50 + Adam + 0.001 + Attention + Transfer Learning	**0.907992**	**0.958873**	**0.795948**	**0.846896**

In Fig. 14 to Fig. 17, it can also be seen that although the attention mechanism yields better results than those that do not use, there is unstable accuracy at epochs 6 and 9. Nevertheless, the results of this study indicate that the attention mechanism can increase the model's accuracy. Therefore, the continuity of this research will use the attention mechanism.

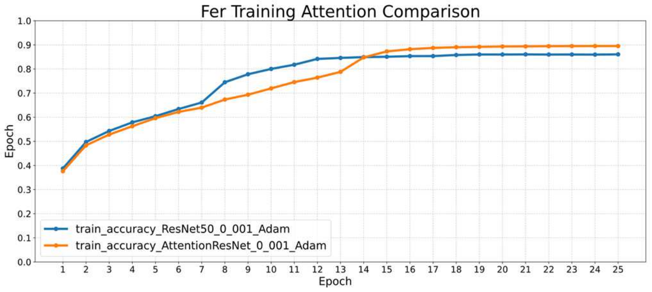

Fig. 14 Effect of Attention Module on Training Data of FER

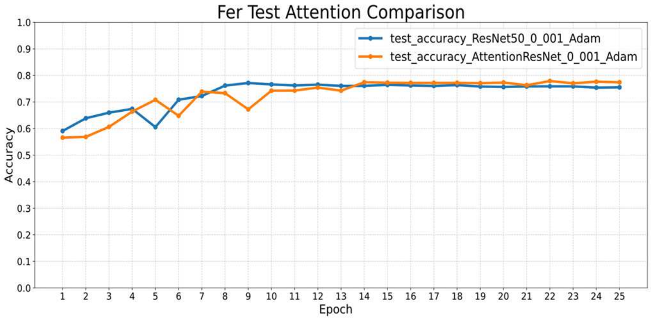

Fig. 15 Effect of Attention Module on Testing Data of FER

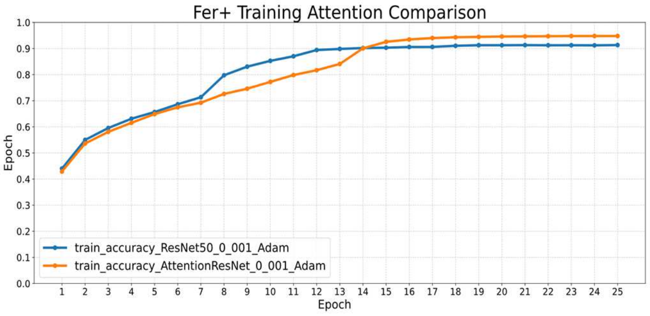

Fig. 16 Effect of Attention Module on Training Data of FER+

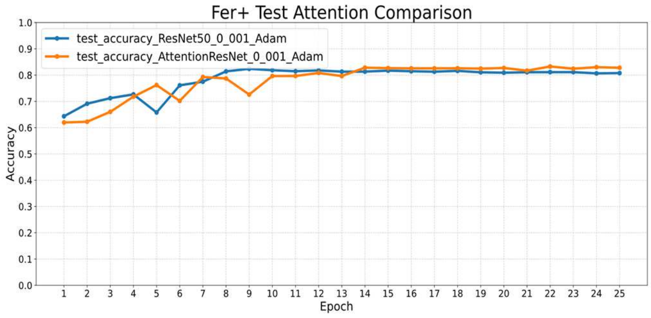

Fig. 17 Effect of Attention Module on Testing Data of FER+

E. Effect of Transfer Learning

In this study, the transfer learning method is utilized by leveraging the weights that have been carried out in previous training (ImageNet). Using the best model attached with attention with a learning rate that has been obtained from previous experiments, the results of using the attention mechanism can be seen in Table IV. All accuracies on the training and test data on the FER and FER+ datasets show that the transfer learning method gives better results than without on model with attention module. It can be seen in Fig. 18 to Fig. 21 that the attention method causes unstable training on the model without transfer learning, which can be seen in the sixth and ninth epochs. However, using the transfer learning method causes the training in the model to be stable in both epochs.

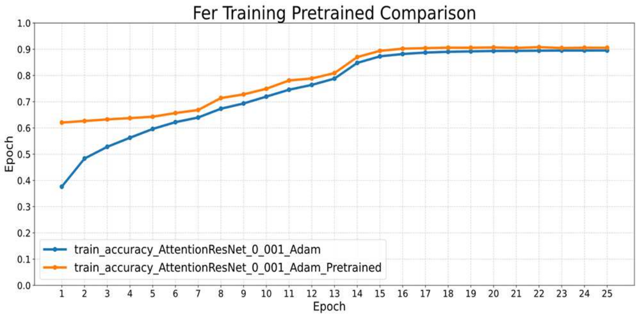

Fig. 18 Effect of Transfer Learning on Training Data of FER

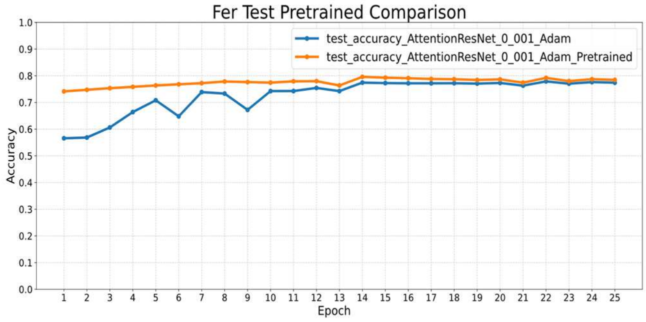

Fig. 19 Effect of Transfer Learning on Testing Data of FER

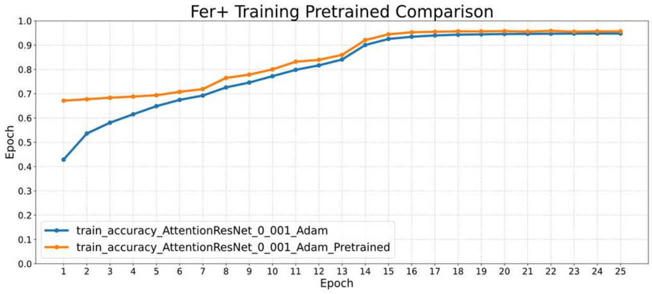

Fig. 20 Effect of Transfer Learning on Training Data of FER+

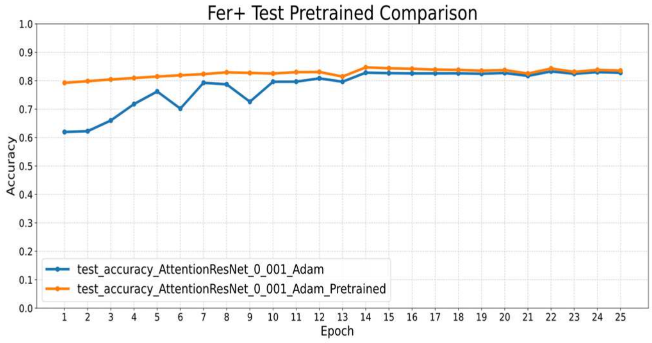

Fig. 21 Effect of Transfer Learning on Testing Data of FER+

F. Classification Error

In this study, we use a confusion matrix to visualize the prediction model's performance. We have calculated for True Positive Rate (TPR) of each dataset's dominant class, with the result for the happy class of FER, and a neutral class of FER+ being 83 % and 81.4%, respectively. While False Positive Rate (FPR) results for the happy class of FER and the neutral class of FER+ are 22.67 % and 25.3%, respectively. However, as shown in the overall confusion matrix of Fig.22 for FER and Fig.23 for FER+, the minority classes of FER+ (angry and disgust) and FER (unknown, fear, disgust, contempt, anger, and NF) show a low TPR and high FPR. Even though overall accuracies are high due to the dominant classes, it can be concluded that some special balancing treatments are required in future works to train imbalanced datasets.

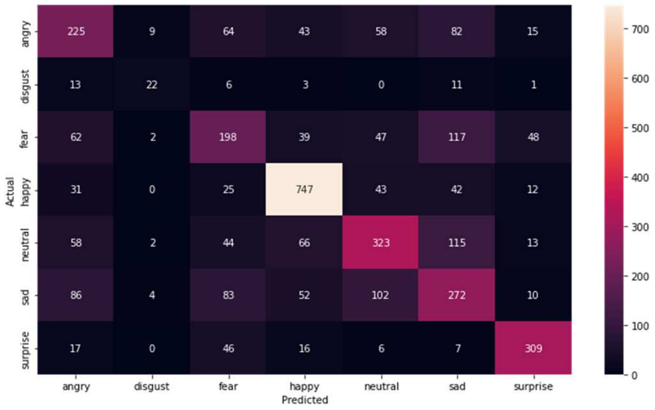

Fig. 22 Confusion Matrix of FER

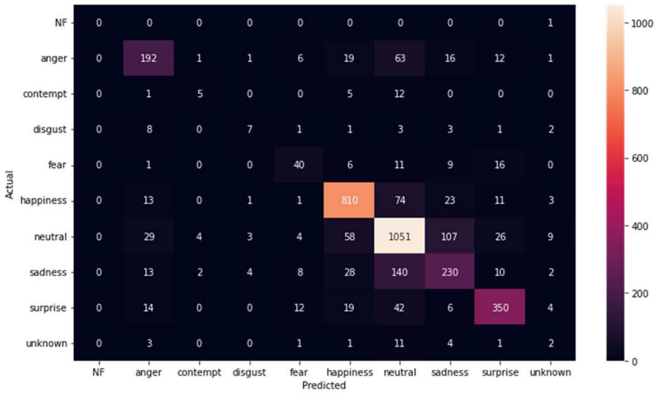

Fig. 23 Confusion Matrix of FER+

We suggest adding more datasets. This study only experiments on two datasets, namely FER and FER+. With the additional datasets, researchers can generalize the best model for facial expression recognition. Moreover, utilizing a scheduler to reduce the learning rate during training can be another option to be investigated. Furthermore, this study shows that there are only six learning rates tested; this allows the opportunity to increase accuracy by using a dynamic and diverse learning rate.

IV. Conclusion

The results of the best accuracy of testing data are ResNet50 model of 0.771539 for FER data and 0.823987 on FER+ data, while for VGGFace are 0.760199 and 0.812477 for FER and FER+ data, respectively. The worst accuracy is produced by baseline CNN, with the results of 0.755404 and 0.809351, for FER and FER+ data, respectively, using hyperparameters of Adam optimizer, with a learning rate of 0.001. The result of using the attention mechanism in this study is increased performance in all the training and testing results with an accuracy of 0.778905 on FER test data and 0.832741 on FER+ data. It concludes that the Attention module is essential in recognizing Facial Expressions using a Convolutional Neural Network (CNN). However, unstable learning during training occurs in the sixth and ninth epochs. Advice for further research is first, adding more datasets besides FER and FER+, and second, adding a Scheduler to decrease the learning rate during the training data. Moreover, weighting mechanisms and other special treatments are candidates as extension methods to train imbalanced datasets.

Acknowledgment

We thank Microsoft for sharing the dataset publicly.

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," 2016, [Online]. Available: http://arxiv.org/abs/1607.06450.

[2] F. Y. Rahadika, N. Yudistira, and Y. A. Sari, "Facial Expression Recognition using Residual Convnet with Image Augmentations," *J. Ilmu Komput. dan Inf.*, vol. 14, no. 2, pp. 127–135, 2021, doi: 10.21609/jiki.v14i2.968.

[3] M. Pourmirzaei, G. A. Montazer, and F. Esmaili, "Using Self-Supervised Auxiliary Tasks to Improve Fine-Grained Facial Representation," 2021.

[4] N. Thseen, "Face-To-Face Communication, Non-Verbal Body Language and Phubbing: the Intrusion in the Process," *Russ. J. Educ. Psychol.*, vol. 11, no. 2, p. 22, 2020, doi: 10.12731/2658-4034-2020-2-22-31.

[5] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychol. Sci. Public Interes.*, vol. 20, no. 1, pp. 1–68, 2019, doi: 10.1177/1529100619832930.

[6] F. Psychol, "Facial Expressions in Context: Electrophysiological Correlates of the Emotional Congruency of Facial Expressions and Background Scenes," *frontiersin.org*, 2017. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02175/full.

[7] S. Turabzadeh, H. Meng, R. Swash, M. Pleva, and J. Juhar, "Facial Expression Emotion Detection for Real-Time Embedded Systems," *Technologies*, vol. 6, no. 1, p. 17, 2018, doi: 10.3390/technologies6010017.

[8] M. J. Taylor, C. Shikaislami, C. McNicholas, D. Taylor, J. Reed, and I. Vlaev, "Using virtual worlds as a platform for collaborative meetings in healthcare: A feasibility study," *BMC Health Serv. Res.*, vol. 20, no. 1, pp. 1–10, 2020, doi: 10.1186/s12913-020-05290-7.

[9] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 3285–3294, 2019, doi: 10.1109/ICCV.2019.00338.

[10] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Trans. Image Process.*, vol. 29, no. 8, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[11] H. Zhou *et al.*, "Exploring emotion features and fusion strategies for audio-video emotion recognition," *ICMI 2019 - Proc. 2019 Int. Conf. Multimodal Interact.*, pp. 562–566, 2019, doi: 10.1145/3340555.3355713.

[12] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," *ICMI 2016 - Proc. 18th ACM Int. Conf. Multimodal Interact.*, pp. 279–283, 2016, doi: 10.1145/2993148.2993165.

[13] T. H. Vo, G. S. Lee, H. J. Yang, and S. H. Kim, "Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020, doi: 10.1109/ACCESS.2020.3010018.

[14] D. Gera, G. N. Vikas, and S. Balasubramanian, *Handling ambiguous annotations for facial expression recognition in the wild*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[15] A. Anton, N. F. Nissa, A. Janiati, N. Cahya, and P. Astuti, "Application of Deep Learning Using Convolutional Neural Network (CNN) Method For Women's Skin Classification," *Sci. J. Informatics*, vol. 8, no. 1, pp. 144–153, 2021, doi: 10.15294/sji.v8i1.26888.

[16] R. Salakhutdinov and G. Hinton, "Replicated softmax: An undirected topic model," *Adv. Neural Inf. Process. Syst. 22 - Proc. 2009 Conf.*, pp. 1607–1614, 2009.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[18] B. Mandal, A. Okeukwu, and Y. Theis, "Masked Face Recognition using ResNet-50," 2021, [Online]. Available: http://arxiv.org/abs/2104.08997.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 2017–2025, 2015.

[20] C. Luna-Jiménez, J. Cristóbal-Martín, R. Kleinlein, M. Gil-Martín, J. M. Moya, and F. Fernández-Martínez, "Guided spatial transformers for facial expression recognition," *Appl. Sci.*, vol. 11, no. 16, 2021, doi: 10.3390/app11167217.