Comparative Analysis of ResNet18 and MaxVit on FER2013 Dataset

Emre Cakmakyurdu
Middle East Technical University
Ankara, Turkey
emre.cakmakyurdu@metu.edu.tr

***Abstract*—Facial expression recognition is important for human-computer interaction as it aids in inferring what the emotions of a human are and how machines react to them. This paper presents an experimental study comparing the performance of two deep learning models, ResNet18 and MaxViT, on the FER2013 dataset. ResNet18, a convolutional neural network, and MaxViT, a hybrid model combining convolutional and transformer-based architectures, were trained and evaluated for their accuracy in classifying facial expressions. The ResNet18 model achieved a test accuracy of 62.4%, while the MaxViT model achieved a test accuracy of 70.58%. The results demonstrate that the MaxViT model outperforms ResNet18, benefiting from its advanced attention mechanisms that capture both local and global features more effectively. These findings suggest that integrating convolutional and transformer-based approaches can significantly enhance facial expression recognition tasks.**

I. Introduction

In human communication, facial expression and body language can provide nonverbal information that can provide additional clues and meanings in verbal communication. There are lots of works conducted to enhance this nonverbal communication into machines. One of the works conducted is facial expression recognition. Facial expression recognition is crucial aspect in human-computer interaction. This problem enhances the way that machines understands and react to human emotions. There are lots of areas in which understanding human emotions plays an important role such as psychology, security, marketing, autopilot and entertainment. In this project, facial expression recognition problem will be tackled using ResNet18 and MaxVit to achieve maximum accuracy using FER2013 dataset.

II. Dataset

FER2013 dataset is introduced in the International Conference in Machine Learning 2013 for a competition which aims to conduct facial expression recognition. The data was collected by scanning over the internet and extracting face images ensuring diverse demographic representations using Google Image Search API. Dataset contains lots of different age, ethnic and background groups. Images were labeled by experts ensuring that each image was annotated with the perceived emotional expression.

Dataset contains 35887 gray-scale images. Each images has a resolution of 48x48. There are seven labels which are angry, disgust, fear, happy, neutral, sad and surprise. The distribution of classes is balanced except for disgust class. Distribution of each class with respect to overall dataset is: Angry: 13.8%, Disgust: 1.5%, Fear: 14.5%, Happy: 25%, Neutral: 17.2%, Sad: 16.9%, Surprise: 11.1%. Dataset is divided as 80% train, 10% validation and 10% test.

The human accuracy on this dataset is 65.5% [1]. Compared to the other datasets, FER has more variation in the images, including face occlusion (mostly with hand), partial faces, low-contrast images, and eyeglasses.

III. Literature Review

A. Facial Expression Recognition using Convolutional Neural Networks: State of the Art

First paper [1] uses FER2013 dataset. Authors applied preprocessing steps such as normalization to have zero mean and standard deviation of 1 and illumination correction such as histogram equalization and linear plane fitting to eliminate the lignting variations effect. Authors discussed several CNN architectures with respect to FER2013 dataset performance. They have tried 6 different shallow CNN's and 3 deep CNN's such as VGG, Inception and ResNet. They tweaked the VGG, Inception and ResNet architectures. They removed one Convolution-Convolution-Pooling layer from VGG and apply dropout after every CCP layers, removed initial strided convolutions or pooling layeres from Inception and removed initial Convolution-Pooling layer from 34 layer ResNet. They concluded that shallower networks performs better than the deeper networks and highlighted the major bottlenecks in CNN-based FER. They showed that by overcoming these bottlenecks and creating an ensemble model approach, modern deep CNN approaches perform better on the FER2013 dataset. Authors train every architecture for up to 300 epochs, optimizing the cross-entropy loss using stochastic gradient descent with a momentum of 0.9. The initial learning rate, batch size, and weight decay are fixed at 0.1, 128, and 0.0001, respectively. The learning rate is halved if the validation accuracy does not improve for 10 epochs. Authors used accuracy as metric. Authors achieved 75.2% accuracy on FER2013 dataset. The baseline in this study is the performance of previous methods using simpler or shallower CNN architectures. There is no code base for the paper.

B. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network

Second paper [2] uses different dataset one of which is FER2013. Authors did not apply any preprocessing steps. The authors propose using an attentional convolutional network that focuses on salient parts of the face crucial for recognizing facial expressions. This approach leverages fewer layers (less than ten) compared to deeper models, yet achieves high performance by concentrating on important facial features rather than the entire image. The proposed model can be found in Figure 1. Each model is trained for 500 epochs from scratch, on an AWS EC2 instance with a Nvidia Tesla K80 GPU. They initialize the network weights with random Gaussian variables with zero mean and 0.05 standard deviation. For optimization, they used Adam optimizer with a learning rate of 0.005 with weight decay. Authors used Adam optimizer with stochastic gradient descent approach. The authors showed three models for baseline. Bag of Words, VGG + SVM and GoogleNet with accuracies 67.4%, 66.31% and 65.2% respectively. They claimed that their model has 70.02% classification accuracy which improves the baseline models. Also, authors proposed a visualization method to highlight the salient regions of face images which are the most crucial parts thereof in detecting different facial expressions. The model does not have any codebase.

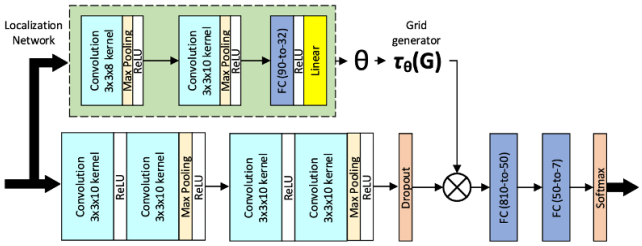

Fig. 1. Model used in second paper.

C. Real-time Convolutional Neural Networks for Emotion and Gender Classification

Third paper [3] focuses on real-time emotion recognition. Authors proposed a new model which can be found in Figure 2. Authors focused on memory usage and inference speed rather than accuracy. They managed to decrease memory usage and increase inference speed without totally decreasing the accuracy. Authors integrates depth-wise seperable convolutinos and residual modules to minimize parameter count. They are inspired by the Xception architecture. The model is trained with Adam optimizer. Paper also uses FER2013 dataset. There is no baseline reported in the paper and authors claimed that they achieved 66% accuracy on the FER2013 dataset. Codebase is available, it can be found in this link.

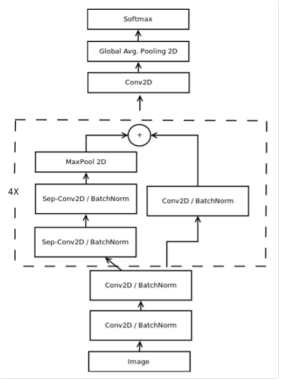

Fig. 2. Model used in third paper.

IV. Methodology

A. Preprocessing

Before training the models, several preprocessing steps were applied to the FER2013 dataset to enhance the quality and consistency of the input images. The steps included:

- **Normalization**: Each image was normalized by dividing the mean and standard deviation by 255.0 to scale the pixel values to the range [0, 1]. This step helps stabilize and speed up the training process by ensuring that the inputs to the network have a consistent distribution.
- **Data Augmentation**: For the training dataset, random horizontal flipping was applied to augment the data. This technique helps prevent overfitting by increasing the variability of the training data.
- **ToTensor Transformation**: All images were converted to PyTorch tensors, which is necessary for processing by the PyTorch neural network models.
- **Resizing**: The images were resized to 224x224 pixels to match the input size required by the MaxViT model. (This step is done only for MaxVit)
- **Gray Scale to RGB**: Since the MaxViT model expects three-channel (RGB) images, the grayscale images from the FER2013 dataset were converted to RGB. (This step is done only for MaxVit).

B. ResNet18

The model presented here, ResNet18, is a deep architecture for the convolutional neural network that was proposed to solve the vanishing gradient problem in building deep networks through the use of residual connections. The ResNet18 architecture is an 18-layered one, and for this case, specific mention is made of both the convolutional layers and residual blocks. The architecture is depicted in 3. Paper can be found in [5]

The first layer is a 7x7 convolution stride two, followed by batch normalization and ReLu activation. The images for this data set are one channel, so the first layer input is adjusted to this. The center of ResNet18 is four residual blocks. Every block has two 3x3 convolutional layers with batch normalization and ReLu activation. Each block has the

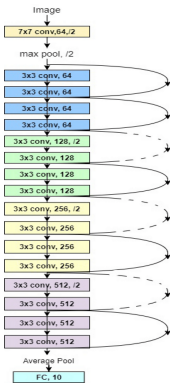

Fig. 3. ResNet18 architecture.

addition of its output to its input through a residual connection. The last layer has been adapted to output the seven emotion classes in the FER2013 dataset.

C. MaxVit

MaxVit is a hybrid architecture that contains, in equal portions, the best benefits from the CNNs and transformers. The vision transformer captures local and global visual parts through the multi-axial attention mechanism. The first layers of MaxVit contain CNN blocks, capturing regional features effectively. Transformer blocks will follow CNN blocks to capture global dependencies in images. The other principal merit of using MaxVit is the multi-axial attention mechanism, which applies to both local and axial attention. It ncaptures dependencies along one dimension and the height and width axes of feature maps independently, thus efficiently acting on long-range dependencies. On the other hand, local attention captures the dependencies within a small region to make sure the details of the images are kept in place. In this paper, ImageNet1k trained version of the model is used to make sure that the model will be able to genneralize well and will be robust enough. The model fits on a GPU with the help of the paper [4].

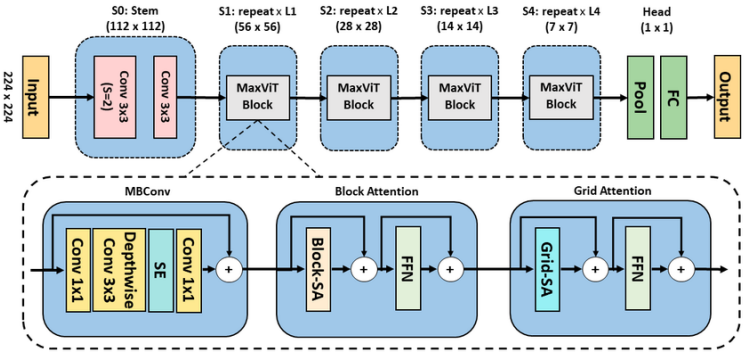

Fig. 4. MaxVit architecture.

MaxVit architecture can be divided into several key structures.

- **Stem Layer**: This is the initial part of the network which processes the input image. It consists two 3x3 convolutional layers with stride 2, which is responsible for reducing the spatial dimensions by half and generating initial feature maps.
- **MBConv Block**: This is an inverted residual block with a squeeze-and-excitation (SE) module. It captures local dependencies and performs depth-wise separable convolutions. 1x1 convolution is responsible for reducing the dimensionality. Depthwise 3x3 convolution is responsible for applying the filter. SE module is responsible for scaling the features. Another 1x1 convolution is responsible for restoring the dimensionality. Also, skip connection adds the input to the output.
- **Block Attention**: This module applies self-attention within each block to capture local context. It consists of multi-head self-attention and feed-forward network (FFN).
- **Grid Attention**: This module applies self-attention across different blocks (grids) to capture global context.
- **Hierarchical Structure**: There are four MaxVit blocks (S1 to S4), each containing a series of MaxViT blocks. The spatial resolution is reduced at each stage while the number of channels increases.
- **Pooling and Fully Connected Layer**: After the final MaxViT block, global average pooling reduces the feature map to a single vector. A fully connected layer maps the pooled features to the output classes.

D. Training

Two models are trained with different procedures. In the following section, details of the training procedures will be explained.

- **ResNet18**: ResNet18 is trained from scratch. No pre-trained weights are used. Adam with a learning rate scheduler based on the validation accuracy is used as the optimizer. The model was trained for up to 100 epochs, with early stopping applied if the validation accuracy did not improve for 10 consecutive epochs. After each epoch, the model was evaluated on the validation set, and the learning rate scheduler was updated based on validation accuracy. If the validation accuracy does not improve for 5 epoch, learning rate is decreased by factor of 0.2. The best model was saved, and final performance was evaluated on the test set. Accuracy is used as main performance metric.
- **MaxVit**: MaxVit is trained using pretrained weights which is taken from hugging face. The final fully connected layer was modified to output the seven emotion classes in the FER2013 dataset. AdamW with a learning rate scheduler based on the validation accuracy is used as the optimizer. he model was trained for up to 50 epochs, with early stopping applied if the validation accuracy did not improve for 6 consecutive epochs. After each epoch, the model was evaluated on the validation set, and the learning rate scheduler was updated based on

validation accuracy. If the validation accuracy does not improve for 10 epoch, learning rate is decreased by factor of 0.2. The best model was saved, and final performance was evaluated on the test set. Accuracy is used as main performance metric.

V. Results and Discussion

A. Results

The best ResNet18 model was trained for 100 epochs with a batch size of 32, learning rate of 0.001, and weight decay of 0. Hyper-parameters are tried using wandb. Learning rate and weight decay searched for 0.001, 0.0001, 0.005 and 0, 0.01 respectively. The highest training accuracy achieved was approximately 99%. The highest validation accuracy achieved was approximately 61%. The final test accuracy was approximately 63%. The training loss decreased steadily, indicating effective learning. The validation loss also decreased, although it showed some fluctuations, indicating the need for early stopping. Results for each hyper-parameter can be found in Table I. Confusion matrix results for ResNet18 can be found in Figure 5. WANDB report can be found in this link.

TABLE I
ResNet18 Results for Different Hyper-parameters

ResNet18	***Test Accuracy***
lr=0.001, wd=0	62.39%
lr=0.001, wd=0.01	0.5681%
lr=0.0001, wd=0	55.2%
lr=0.0001, wd=0.01	55.78%
lr=0.005, wd=0	61.1%
lr=0.005, wd=0.01	45%

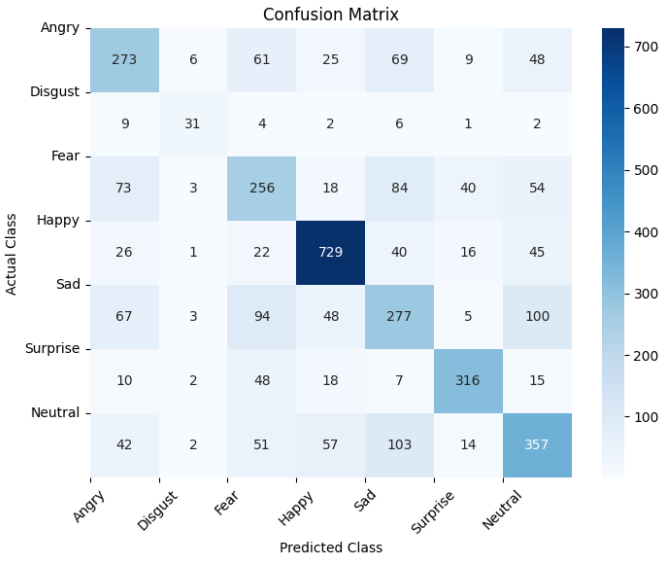

Fig. 5. ResNet18 Confusion Matrix.

The best MaxVit model was trained for 50 epochs with a batch size of 32, learning rate of 0.0005, and weight decay of 0.001. Hyper-parameters are tried using wandb. Learning rate and weight decay searched for 0.0001, 0.0005 and 0, 0.001 respectively. The highest training accuracy achieved was approximately 96%. The highest validation accuracy achieved was approximately 70%. The final test accuracy was approximately 70.58%. The training loss decreased steadily, indicating effective learning. The validation loss also decreased, although it showed some fluctuations, indicating the need for regularization techniques. Results for each hyper-parameter can be found in Table II. Confusion matrix results for MaxVit can be found in Figure 6. WANDB report can be found in this link.

The confusion matrix shows that the MaxViT model shows that model performs well in recognizing the 'Happy' and 'Neutral' classes, and slightly better in recognizing other classes compared to ResNet18.

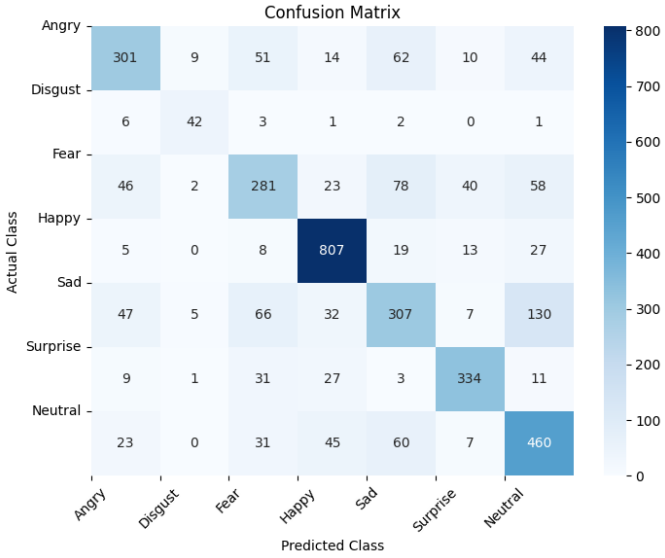

Fig. 6. MaxVit Confusion Matrix.

TABLE II
MaxVit Results for Different Hyper-parameters

MaxVit	***Test Accuracy***
lr=0.0001, wd=0	70.52%
lr=0.0001, wd=0.001	70.58%
lr=0.0005, wd=0	68.1%
lr=0.0005, wd=0.001	69%

B. Discussion

Both of the model results showed that this MaxViT model was more accurate and generalizing towards the test set compared to the ResNet18 model. Generally, MaxViT attains better performance as compared to the ResNet18 model on validation and test sets. Thus, it can be concluded that the MaxViT architecture contains both the convolutional and self-attention mechanisms, which are better at capturing local features and generalizing important global attributes of facial expression images. Both models indicated reducing training and validation losses in the learning curves. The trend of

loss values can be seen as a bit stabilized for the MaxViT model, possibly due to the advanced attention mechanisms to learn and generalize better. From the confusion matrices, both models perform well in classes 'Happy' and 'Neutral.' Detection of the other classes is usually pretty tricky because of slightly varied features. Large values of convergence, in this regard, showed that the ResNet18 model was best compared to the MaxViT model. Both these models performed well, not letting overfitting occur through early stopping features. Where ResNet18 is relatively more straightforward and can be easily trained, the MaxViT-trained model is more competent because it can capture all the fine-grained patterns with the self-attention mechanisms.

VI. Future Work

This study shows that hybrid architectures can be a good candidate for facial expression recognition. However, implementing more sophisticated data augmentation techniques such as CutMix, MixUp, and RandAugment can be done to increase test set accuracy.

VII. Conclusion

ResNet18 and MaxViT architectures are compared for the study regarding the recognition of facial expression.The goal was to classify face images according to the cross entropy loss to identify the emotion. ResNet18, the model used as a baseline, yielded 62% accuracy on classification for test sets. This performance further demonstrates the effectiveness of CNNs in capturing local features and patterns within facial images. ResNet18 is robust and widely chosen for most image classification tasks. Its architecture includes residual connections, enabling the training of deeper networks without facing the vanishing gradient problem. Finally, to further enhance the performance, fine-tuned pre-trained model with the ImageNet1k dataset on the MaxViT model is trained. The test accuracy of the MaxViT model is 70.52% on the test set, which is even better than that of ResNet18. The superior performance of MaxViT is attributed to its hybrid architecture that fuses CNNs and transformers. As a result, the initial layers in the MaxViT have convolutions, which help in effectively capturing local features. In contrast, the transformer layers will capture long-range dependencies and global context in the image. In a word, while ResNet18 showed good performance, the MaxViT model outperformed the model of ResNet18 in the task of facial expression recognition.

References

[1] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," arXiv preprint arXiv:1612.02903, December 2016.

[2] S. Minaee and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," arXiv preprint arXiv:1902.01019, February 2019.

[3] O. Arriaga, P. G. Plöger, and M. Valdenegro, "Real-time Convolutional Neural Networks for Emotion and Gender Classification," arXiv preprint arXiv:1710.07557, October 2017.

[4] M. T. Hassani, A. Walton, J. Navon, A. N. Ford, T. Aila, and A. Gholami, "MaxViT: Multi-Axis Vision Transformer," arXiv preprint arXiv:2204.01697, 2022.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," arXiv preprint arXiv:1512.03385, 2015.