An Extensive Investigation of Deep Learning Models for Recognizing Face Emotions

Arun Uday Chaudhary,[a] Siddharth,[b] and Dr. Chhavi Dhiman[c]

Department of Electronics and Communication Engineering, Delhi Technological University, Shahbad Daulatpur, 110042 Delhi, India.

[a]Corresponding author: arunudaykvic@gmail.com

[b]siddharthsingh25102001@gmail.com

[c]chhavi.dhiman@dtu.ac.in

Abstract. Facial Emotion Recognition (FER) is an important field with applications in security, healthcare, and human-computer interaction, among others. Because deep learning models can automatically learn hierarchical representations from raw data, they have demonstrated promising outcomes in FER challenges. Using the FER2013 dataset, we give a comparative analysis of three popular Deep Learning architectures for FER: ResNet50, Vision Transformer (ViT), and VGG16. The main goal is to assess and contrast these models' abilities to identify facial expressions. We trained each model for 50 epochs due to restricted computing resources. We carry out in-depth studies taking into account a number of variables, including accuracy, computing efficiency, and model complexity. We also look at how transfer learning and data augmentation methods affect model performance. Our results shed light on the advantages and disadvantages of each design with regard to FER tasks. This research attempts to offer useful information for choosing suitable Deep Learning models for face emotion identification applications through thorough examination and comparison.

Keywords: Facial Emotion Recognition (FER), human-computer interaction, healthcare, security, deep learning models, hierarchical representations, raw data, comparative analysis, ResNet50, Vision Transformer (ViT), VGG16, FER2013 dataset, model complexity, computational efficiency, accuracy, data augmentation techniques, transfer learning, performance evaluation, model strengths and limitations, deep learning architectures.

INTRODUCTION

An important field in human-computer interaction is Facial Emotion Recognition (FER), which allows computers to deduce emotions from facial expressions. The effectiveness of three deep learning models—ResNet50, VGG16, and Vision Transformer (VIT)—in identifying face emotions is examined in this study. We created custom layers and refined these models on the FER2013 dataset using pretrained versions of these models. Owing to limitations in processing power, every model underwent 50 training cycles. In order to demonstrate these models' possible uses in real-world situations where a knowledge of human emotions is crucial, this study compares the resilience and performance of the models.

BACKGROUND STUDY

Introduction to Facial Emotion Recognition

Using computational techniques, Facial Emotion Recognition (FER) recognizes human emotions from facial expressions. Improving human-computer interactions through the ability of robots to react correctly to human emotions is crucial. With a great improvement in accuracy and dependability, technology has progressed from basic image processing methods to sophisticated machine learning and deep learning algorithms.

Applications of FER

Applications for FER are widespread and include optimizing the user experience in HCI, enhancing security via behavior analysis, supporting mental health diagnosis and treatment, personalizing marketing campaigns via consumer emotion analysis, and enabling emotionally intelligent robots and virtual assistants.

Description of the FER2013 Dataset

The 35,887 grayscale pictures of faces in the FER2013 dataset were produced for the ICML 2013 Challenges in Representation Learning. Each image has a label associated with one of seven emotion categories. It is a common

benchmark for FER system evaluation and training due to its size and diversity.

Pretrained Deep Learning Models

Neural networks with extensive dataset training are used to create pre-trained models such as ResNet50, VGG16, and Vision Transformer (VIT). VGG16 is renowned for its depth and simplicity, ResNet50 for its residual learning framework, and VIT for using transformer architecture in vision tasks. These models are good starting points for additional fine-tuning on particular tasks such as FER.

Custom Layer Addition and Fine-Tuning

Modifying the architecture of pretrained models to better fit the FER job is known as adding custom layers, and this usually entails adding layers that capture emotion-specific data. By modifying the pretrained model's weights on the fresh dataset, fine-tuning improves performance and modifies the model to fit the unique nuances of the FER2013 facial expressions.

Related Works

The performance characteristics of facial expression recognition (FER) utilizing machine learning models have been investigated in a number of earlier works. A thorough overview of automatic facial affect analysis was presented by Sariyanidi, Gunes, and Cavallaro [1], with an emphasis on registration, representation, and recognition methods. Martinez [2] outlined the developments, difficulties, and prospects in automatic FER. Goodfellow et al.'s [3] report on representation learning issues was based on three machine learning competitions. In a seminal study that has impacted numerous FER systems, Krizhevsky, Sutskever, and Hinton [4] proved the effectiveness of deep convolutional neural networks for ImageNet classification. Deep residual learning was introduced by He et al. [5], greatly enhancing picture recognition tasks. FaceNet, a unified embedding for face identification and clustering, was proposed by Schroff, Kalenichenko, and Philbin [6] and shown excellent face verification accuracy. Tang [7] examined support vector machines for deep learning, whereas Yu and Zhang [8] used multiple deep network learning for static facial expression identification. Aligned and non-aligned face information were integrated by Kim et al. [9] to improve FER in practical settings. Through the EmotiW 2015 challenge, Dhall et al. [10] explored the difficulties associated with image- and video-based emotion identification in the outdoors. In order to further push the bounds of accuracy and reliability, Kim et al. [11] have introduced a hierarchical committee of deep convolutional neural networks for resilient FER. Using cutting-edge deep neural networks, Mollahosseini, Chan, and Mahoor [12] investigated FER in further detail.

METHODOLOGY

This section describes the methodology used to assess the efficacy of three deep learning models for Facial Emotion Recognition (FER) using the FER2013 dataset: ResNet50, VGG16, and Vision Transformer (VIT). To guarantee repeatability and a reliable performance comparison, the technique consists of the following steps: preparation of the dataset, model customisation, transfer learning and fine-tuning, training protocols, evaluation metrics, and experimental setup.

Dataset

A widely used publicly accessible dataset for facial expression recognition applications is FER2013. One of the following seven emotions is assigned to each image in the dataset: surprise, happiness, sorrow, disgust, anger, and neutral. In order to meet the models' input requirements, preprocessing entailed scaling the photos to a standard resolution of 224x224 pixels. The training pictures underwent data augmentation procedures to mimic a more varied dataset and avoid overfitting. These methods included horizontal image flipping, filling mode, random rotations, and zooming in and out. For the purposes of training and validation, we employed an 80-20 split. We also assigned class weights to the respective 7 classes in the FER2013 dataset to address the issue of class imbalance in the FER2013 dataset.

Classes	Number of images
Angry	2466
Disgust	191
Fear	652
Happy	7528
Neutral	10300
Sad	3514
Surprise	3562
Contempt	165

FIGURE 1: Training Dataset

Classes	Number of images
Angry	644
Disgust	57
Fear	167
Happy	1827
Neutral	2597
Sad	856
Surprise	900
Contempt	51

FIGURE 2: Test Dataset

FIGURE 3: FER2013 Dataset

Pre-trained models

Pre-trained models serve as a strong foundation due to their learned feature extraction capabilities from the extensive ImageNet dataset. For each model:

- ResNet50: Renowned for its foundation of residual learning. After the convolutional basis of ResNet50, other layers were added. These included a Flatten layer for reshaping, two Dense layers with ReLU activation, BatchNormalization, and Dropout for regularization, after each other. In order to facilitate classification, a Dense layer with softmax activation was attached.
- VGG16: This model's simplicity and depth make it suitable for extracting hierarchical features. Similar custom layers were added post the convolutional base.
- Vision Transformer (VIT): VIT leverages transformer architecture for image recognition. Custom layers included dense layers after the transformer block to adapt it to the FER task.

Transfer Learning and Fine-Tuning

Transfer learning allows leveraging the knowledge embedded in pretrained models. The three models were pre-trained on the 'Imagenet' dataset . Initial layers were frozen to retain the generic features, while custom layers were unfrozen to fine-tune the model to FER-specific features. A low learning rate (e.g., 0.0001) was used during fine-tuning to avoid large updates that could disrupt the learned features. The Adam optimizer was selected for its adaptive learning rate properties, and the categorical cross-entropy loss function was used to measure prediction error.

Training

Training was conducted in a controlled environment in Google Colab to handle the computational demands. The training process involved splitting the dataset into training, validation, and test sets. The batch size was set to 64 to balance memory usage and training speed. Early stopping was used to monitor validation loss and halt training if no improvement was observed, thereby preventing overfitting. Dropout layers were incorporated to randomly drop units during training, which helps in generalizing the model.

Parameters

The 7 classes mentioned above were trained on identical training subsets and subsequently tested against the same testing data subsets. To ensure a sound comparison, we defined four key parameters based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These parameters include:

- Accuracy measures the proportion of correctly predicted instances out of the total instances.
- Precision indicates the accuracy of positive predictions.
- Loss quantifies the error between the predicted and actual labels.Lower loss values indicate better model performance. Important for understanding how well the model is learning and optimizing over epochs.

- AUC (Area Under the Curve) represents the ability of the model to distinguish between classes. A higher AUC value indicates better performance in terms of classification. Provides insights into the model's overall ability to correctly rank positive instances higher than negative ones.

RESULT

In order to assess the performance of three well-known deep learning models—ResNet50, VGG16, and Vision Transformer (VIT)—we carried out a thorough comparison study of our research on the FER2013 dataset. For both the training and validation sets across a 50-epoch period, we used accuracy, loss, Area Under the Curve (AUC), and precision as our assessment measures. The findings show clear variations in the models' capacities for generalization and learning dynamics:

ResNet50

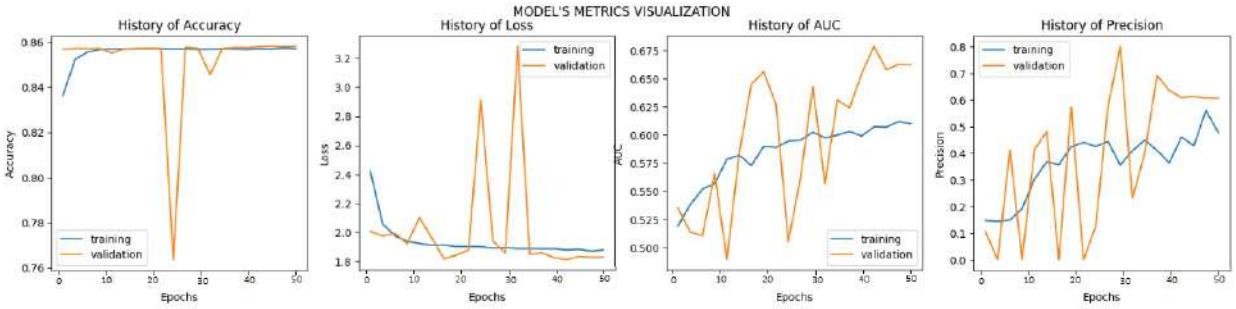

FIGURE 4: Respective Metrics Graph for ResNet50.

Accuracy: Training accuracy is high at first and rapidly settles around 0.86, suggesting that the model does a good job of learning the training set. Similar to the initial high validation accuracy, there are notable declines at epochs 25 and 35, which point to possible overfitting and instability.

Loss: Effective learning is demonstrated by the training loss's steady decline. However, there are many peaks in the validation loss, which indicates instability and overfitting.

AUC: The training AUC steadily rises, indicating that the model's performance in class distinction is becoming better. Because of the considerable volatility of the validation AUC, the validation set performance was found to be inconsistent.

Precision: Better precision on the training set is shown by the training precision's gradual increase. The substantial fluctuations in the validation precision indicate that the model's accuracy varies widely between validation sets.

VGG16

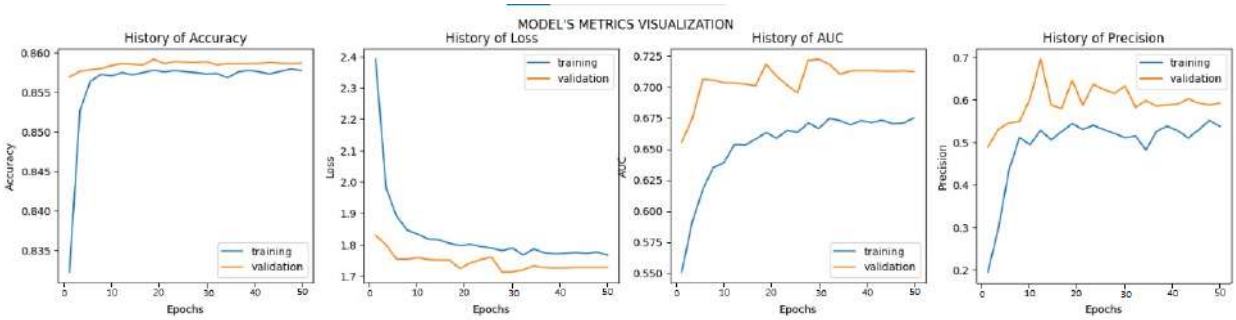

FIGURE 5: Respective Metrics Graph for VGG16.

Accuracy: Training accuracy increases quickly before stabilizing around 0.86, suggesting effective learning. Good generalization and model stability are shown by the validation accuracy, which fluctuates somewhat but stays quite close to the training accuracy.

Loss: Effective learning is shown by the training loss, which first exhibits a considerable decline before stabilizing. Consistently lowering the validation loss relative to the training loss suggests both possible overfitting and strong generalization.

AUC: The training AUC continuously rises, demonstrating a better capacity for class distinction. Consistent performance between the training and validation sets is shown by the validation AUC, which stays comparatively steady and near to the training AUC.

Precision: Over time, the training precision increases, suggesting that the model gets more accurate. With just slight variations, the validation precision is comparatively steady, indicating consistent precision across several validation sets.

VIT Model

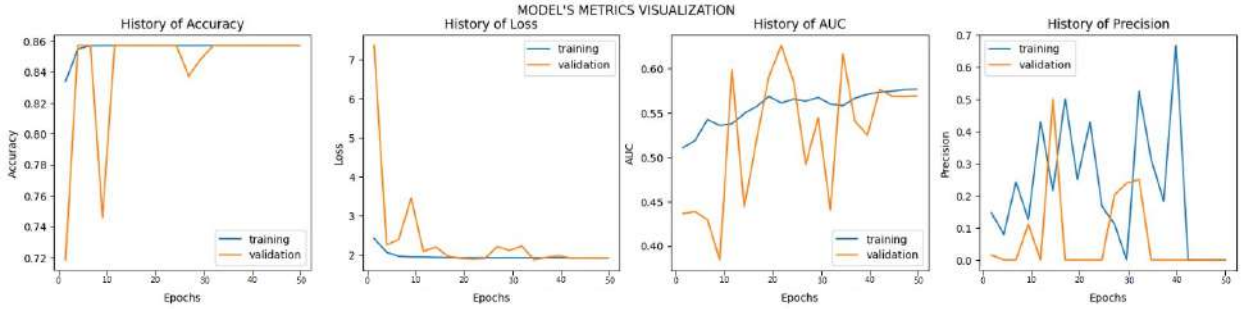

FIGURE 6: Respective Metrics Graph for VIT.

Accuracy: Training accuracy is high at first and stays steady around 0.86, suggesting that learning is taking place. Extreme changes in the validation accuracy point to instability and perhaps overfitting.

Loss: Good learning is shown by the training loss's continuous decline. Significant fluctuations in the validation loss are indicative of instability and perhaps overfitting.

AUC: The training AUC steadily rises, indicating a better capacity for class distinction. High volatility in the validation AUC indicates uneven performance on the validation set.

Precision: Better precision on the training set is indicated by the training precision's steady improvement over time. The significant volatility of the validation accuracy indicates uneven precision between validation sets.

CONCLUSION

VGG16 is the best appropriate model among the three for face emotion recognition because of its reliable and consistent results on both training and validation sets. With just slight overfitting, it exhibits good generalization. Although **ResNet50** and **VIT** perform well on training data, they still need to be further optimized to deal with overfitting and instability. Their performance may be enhanced by using regularization strategies including dropout, data augmentation, and early halting in addition to making sure the validation set is representative and consistent.

Ultimately, **VGG16** is suggested as the most dependable model in this comparison for the identification of facial emotions, whereas **ResNet50** and **VIT** exhibit promise but require further fine-tuning and regularization to attain comparable stability and efficacy.

FUTURE WORKS

A number of focused measures based on ResNet50, VGG16, and Vision Transformer (VIT) performance may be implemented to enhance face emotion identification. In order to stabilize validation performance, it is imperative to address overfitting and instability, especially for ResNet50 and VIT, by utilizing more robust regularization approaches like dropout and L2 regularization. In order to boost data variety and prevent bias towards frequent classes, advanced

data augmentation techniques like random cropping and rotation should be used to all models. Additionally, class balance in the training data should be maintained. Ensemble techniques may be used to lower variability and boost overall performance, and k-fold cross-validation should be used for consistent performance evaluation to increase generalization. In order to improve model design, deeper feature representations can be accessed by fine-tuning models using pretrained weights from bigger datasets, or by using VGG16 as a feature extractor in conjunction with models such as LSTM to capture temporal relationships. Critical parameters such as batch sizes and learning rates should be optimized for all models using automated hyperparameter tuning strategies. Enhancing patch embedding techniques and fine-tuning attention processes will aid VIT in more accurately capturing significant face characteristics. Ultimately, assessing the robustness of the model in a variety of real-world scenarios and using adversarial training can strengthen defenses against possible intrusions and guarantee reliable real-world performance. With these actions, we want to reduce overfitting, enhance generalization, and guarantee accurate face emotion identification in real-world scenarios.

REFERENCES

1. E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE transactions on pattern analysis and machine intelligence **37**, 1113–1133 (2014).
2. B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," Advances in face detection and facial image analysis , 63–100 (2016).
3. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *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* (Springer, 2013) pp. 117–124.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems **25** (2012).
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.
6. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 815–823.
7. Y. Tang, "Deep learning using support vector machines," CoRR, abs/1306.0239 **2** (2013).
8. Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction* (2015) pp. 435–442.
9. B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2016) pp. 48–57.
10. A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on international conference on multimodal interaction* (2015) pp. 423–426.
11. B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," Journal on Multimodal User Interfaces **10**, 173–189 (2016).
12. A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)* (IEEE, 2016) pp. 1–10.