Facial Expression Recognition via Vision Transformer Fine-Tuning

Jiacheng You
Master of Applied Economics
University of Michigan
Email: jiacheny@umich.edu

***Abstract*—Facial expression recognition is vital for enhancing human-computer interaction by interpreting human emotions. This paper fine-tunes a Vision Transformer (ViT) model for classifying facial expressions into seven distinct emotion categories. Utilizing the FER2013 dataset with data augmentation and advanced preprocessing techniques, the model achieves an accuracy of 67.22%. Results show strong performance in detecting ”happy” and ”surprise” emotions but reveal room for improvement in identifying ”fear” and ”sadness”. Future enhancements could focus on additional training data, hyper-parameter tuning, or ensemble methods.**

I. Introduction

Facial expression recognition (FER) plays a crucial role in enhancing human-computer interaction by enabling machines to interpret and respond to human emotions in a natural and intuitive manner. The ability to accurately recognize facial expressions allows systems to better adapt to users' emotional states, improving user experience in applications such as virtual assistants, education, healthcare, and social robotics.

Traditional FER methods often rely on convolutional neural networks (CNNs), which have shown effectiveness in many visual tasks. However, these approaches may struggle with variations in head pose, lighting conditions, and occlusions commonly found in real-world settings. To address these limitations, Vision Transformers (ViTs) have emerged as a powerful alternative. Leveraging self-attention mechanisms, ViTs are capable of modeling long-range dependencies and capturing global contextual information, which is particularly beneficial for expression analysis.

Recent studies have demonstrated the superiority of transformer-based models in FER tasks. Ma et al. showed that attention-based fusion techniques within Vision Transformers can significantly improve emotion classification accuracy by effectively integrating multi-scale facial features [1]. Building upon this, Aouayeb et al. introduced a hybrid architecture that integrates Squeeze-and-Excitation (SE) blocks with ViTs, leading to enhanced feature representation and improved robustness against intra-class variability [2]. Furthermore, Li et al. proposed the Poker Face Vision Transformer (PF-ViT), which incorporates a disentanglement module to isolate emotion-relevant features from identity-related noise, thereby achieving state-of-the-art performance on multiple benchmark datasets [3].

These advancements highlight the growing potential of transformer-based architectures in FER. By exploiting attention mechanisms and feature disentanglement strategies, recent approaches have addressed many of the challenges faced by earlier CNN-based models. This paper builds on this progress by fine-tuning a ViT model on the FER2013 dataset, aiming to further explore the effectiveness of attention-driven architectures in recognizing subtle and complex emotional expressions.

II. Methodology

The problem involves classifying grayscale facial images from the FER2013 dataset into seven emotions: angry, disgust, fear, happy, neutral, sad, and surprise. The preprocessing steps included using Hugging Face's AutoImageProcessor and augmenting images with Gaussian blur and brightness adjustments to enhance robustness.

The Vision Transformer model from Hugging Face (ViTForImageClassification) was fine-tuned with the following specifications:

- Optimizer: AdamW
- Learning Rate: 2e-5
- Loss Function: CrossEntropyLoss

Batch inference methods were employed for efficient training and evaluation.

III. Results

The fine-tuned ViT model achieved an overall accuracy of 67.22%. Performance varied across emotion categories:

- Highest performance: ”happy” (F1-score: 0.86) and ”surprise” (F1-score: 0.78)
- Lower performance: ”fear” (F1-score: 0.52) and ”sad” (F1-score: 0.56)

Overall weighted averages were precision: 0.67, recall: 0.67, and F1-score: 0.67, indicating balanced but moderate model efficacy.

IV. Conclusion

This study successfully demonstrated that fine-tuning a Vision Transformer for facial expression recognition yields moderate accuracy, with specific strengths and weaknesses across emotion categories. Future research directions include expanding training datasets, optimizing hyperparameters, and adopting ensemble approaches to further enhance classification accuracy, particularly for less accurately classified emotions.

REFERENCES

[1] Ma, F., Sun, B., & Li, S. (2021). *Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion*. arXiv preprint arXiv:2103.16854.

[2] Aouayeb, M., Hamidouche, W., Soladie, C., Kpalma, K., & Seguier, R. (2021). *Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition*. arXiv preprint arXiv:2107.03107.

[3] Li, J., Nie, J., Guo, D., Hong, R., & Wang, M. (2022). *Emotion Separation and Recognition from a Facial Expression by Generating the Poker Face with Vision Transformers*. arXiv preprint arXiv:2207.11081.