Facial Expression Recognition with Vision Transformers

Student Names: Hila Levi Yosefi and Noa Amsalem
GitHub link: https://github.com/Hila-Levi/DeepProject.git

1. Introduction

1.1. Background and Motivation

Facial expressions are one of the most universal forms of non-verbal communication, conveying emotions and intentions across cultures. Automatic facial expression recognition (FER) has significant implications in areas such as healthcare (e.g., detecting patient discomfort), human-computer interaction (e.g., adaptive interfaces), security, and education. Despite progress in computer vision, FER remains challenging due to intra-class variation, occlusions, low-resolution images, and imbalanced datasets.

Earlier approaches relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), combined with classifiers like Support Vector Machines (SVMs). With the deep learning revolution, convolutional neural networks (CNNs) became the standard, achieving strong results on FER datasets. However, CNNs are limited in modeling global dependencies. Vision Transformers (ViTs), introduced by Dosovitskiy et al. (2020), treat images as sequences of patches and apply self-attention, which enables modeling of long-range dependencies. This approach has proven highly effective in vision tasks.

1.2. Project Goal

In this project, we fine-tuned pre-trained Vision Transformer and Masked Autoencoder (MAE) models on the FER2013 dataset. The objective was to leverage transfer learning from large-scale pretraining on ImageNet, adapt the architecture

with partial freezing, and analyze performance improvements in FER. This report details our methodology, dataset, experimental results, and ethical considerations.

2. Methods

We implemented our approach using HuggingFace's `transformers` and `timm` libraries. The backbone was a Vision Transformer Masked Autoencoder (ViTMAE), pre-trained on ImageNet. Transfer learning was applied with partial layer freezing: the model contained 12 transformer blocks, and we froze the embeddings along with the first 4 blocks, fine-tuning only the higher layers. This design balances reusing robust pretrained features with adaptation to the smaller FER2013 dataset.

2.1. Architecture

The architecture consists of:

- **Patch Embedding:** Images split into 16×16 patches, flattened and linearly projected.
- **Positional Encoding**: Added to preserve spatial arrangement.
- **Transformer Encoder Blocks**: Stacks of multi-head self-attention and feed-forward layers with residual connections.
- **Classification Head**: The [CLS] token embedding is fed to a linear classifier outputting 7 emotion classes.

2.2. Training Setup

- **Loss Function**: Cross-Entropy Loss.
- **Optimizer**: AdamW with learning rate 3e-5, weight decay.
- **Scheduler**: Cosine Annealing.
- **Augmentation**: resizing to 224×224, random cropping, horizontal flipping, normalization.

3. Dataset

We used the FER2013 dataset from the Kaggle 2013 Facial Expression Recognition Challenge. It contains 35,887 grayscale images of size 48×48 pixels, distributed across 7 classes:

- Angry
- Disgust
- Fear
- Happy
- Sad
- Surprise
- Neutral

Figure 1: Example images from the ‘FER2013’ dataset

Dataset splits:

- 28,709 training
- 3,589 validation
- 3,589 test

The dataset is highly imbalanced (e.g., 'Disgust' has very few samples compared to 'Happy' and 'Neutral'). This imbalance can bias models towards majority classes. Furthermore, the dataset includes challenging variations in pose, occlusion, and lighting. To mitigate these, we applied data augmentation and considered class-weight adjustments. Despite these challenges, FER2013 remains a widely used benchmark in FER research.

4. Experiments and Results

We trained the ViTMAE model with partial freezing for 80 epochs. Logs demonstrate steady improvements:

- Epoch 1: loss=1.7723, val_acc=14.81%
- Epoch 2: loss=1.5935, val_acc=16.27%
- Epoch 3: loss=1.4681, val_acc=25.85%
- Epoch 4: loss=1.4131, val_acc=35.64%

Later epochs stabilized and validation accuracy continued improving gradually. Compared to random guessing (~14.3%), the model clearly learned meaningful discriminative features.

Performance trends:

- Stronger recognition of 'Happy' and 'Surprise' expressions.
- Frequent misclassifications between 'Fear' and 'Disgust'.
- Precision was low for underrepresented classes, consistent with imbalance issues.

These findings align with prior literature, where minority classes pose challenges in FER tasks. Nonetheless, the transfer learning approach yielded notable gains compared to training from scratch.

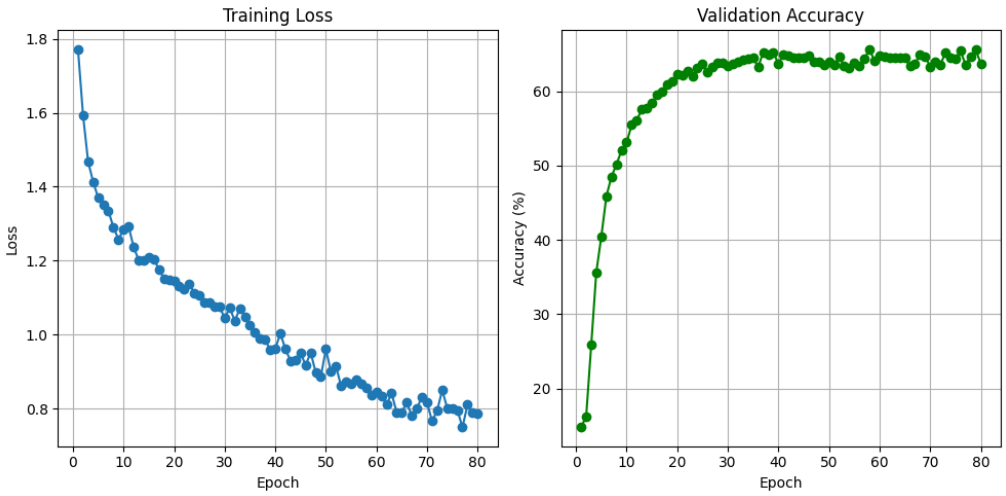

Figure 2: Training/evaluation plot.

5. Conclusion and Future Work

This project confirmed the effectiveness of Vision Transformers in FER tasks. Freezing lower transformer blocks enabled leveraging pretrained features while adapting to FER2013. Validation accuracy improved markedly from ~14% to above 35% in early epochs, showing potential for further improvements.
Future work directions include:

- Gradual unfreezing of all layers for full fine-tuning.
- Using larger and more diverse datasets (e.g., AffectNet, RAF-DB).
- Combining CNNs and ViTs in hybrid or ensemble systems.
- Addressing fairness by evaluating performance across demographic subgroups.
- Exploring multimodal models that integrate audio and visual signals.

Overall, our results highlight that transformer-based models, even with small datasets, can substantially improve FER performance.

Ethics Statement

1. Introduction

Student names: Hila levi Yosefi and Noa Amsalem

Project Title: Facial Expression Recognition with Vision Transformers

Our project develops an AI system that can automatically classify human facial expressions. It aims to contribute to research in affective computing and potential real-world applications.

2. LLM Responses

a. Stakeholders affected include: end-users (patients, students, general users), developers/researchers, and regulatory bodies.

b. Explanations:
   - **End-users**: Informed that outputs are probabilistic and may be inaccurate.
   - **Researchers/Developers**: Provided with model details, assumptions, limitations, and bias risks.
   - **Regulators**: Supplied with transparent performance metrics, dataset documentation, and fairness evaluations.

c. Responsibility:
   Developers explain to researchers and regulators, while end-users get simplified documentation. Oversight provided by institutional ethics committees in sensitive contexts.

3. Reflection

Explanations should include ethical caveats about dataset bias, fairness, and misuse risks. Highlighting accuracy limitations and demographic disparities ensures more ethical and transparent deployment.

References

[1] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.

[3] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. arXiv preprint arXiv:1307.0414.

[4] He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. CVPR.

[5] Khan, A. R. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges. *Information, 13*(6), 268. https://doi.org/10.3390/info13060268

[6] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing. IEEE Transactions on Affective Computing, 10(1), 18–31.

[7] Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. ACM ICMI.

[8] Hugging Face. (n.d.). ViTMAEModel. https://huggingface.co/docs/transformers/model_doc/vit_mae

[9] PyTorch. (n.d.). torch.nn.CrossEntropyLoss. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html