Computer Vision Final Report: Facial Expression Recognition

Student: Manh Ha Nguyen, Le Thuy An Phan

Abstract

This report details our strategy for efficient and accurate facial emotion recognition on the FER-2013 dataset, which faces challenges like high class imbalance and a 65% human accuracy benchmark. Our approach balances performance and computational cost using an adapted, pretrained **EfficientNet-B0** model. Gains over baseline were achieved through specific architectural modifications for efficiency, **advanced data augmentation** (e.g., MixUp), and **optimised multi-stage training**. Crucially, we leveraged **ensemble modelling** of multiple smaller, efficient networks. Our submitted two-model EfficientNet-B0 ensemble achieved **0.6874 accuracy** with a theoretical combined cost of **0.0464 GFLOPs**, demonstrating a strong accuracy-to-cost trade-off.

1. Introduction

Facial Emotion Recognition (FER) on the FER2013 dataset presents challenges due to low-resolution grayscale images and **high class imbalance**. With an **average human accuracy of 65%**, our goal was to develop a system that significantly surpasses this benchmark while maintaining computational efficiency. This report outlines our methodology for achieving these performance gains and optimising the accuracy-cost trade-off.

2. Methodology for Performance & Efficiency Gains

Our core strategy involved optimising a pretrained EfficientNet-B0 model and utilising ensemble techniques.

2.1 Base Model Adaptation and Efficiency

We adapted a **pretrained EfficientNet-B0** as our base. Efficiency gains were primarily achieved by:

- **Reducing Input Resolution:** Instead of the typical 224x224 pixel input expected by standard pretrained models, we adapted the EfficientNet-B0 to process the FER-2013 dataset's native 48x48 pixel resolution, significantly cutting computational load.

- **Single-Channel Input:** Adapted the initial layer for **1-channel grayscale** images, avoiding redundant computations.

- **Streamlined Classifier Head:** Replaced the large 1000-class ImageNet head with a compact **7-class linear layer**. These adaptations resulted in a highly efficient single EfficientNet-B0 model, approximately **0.0232 GFLOPs**.

2.2 Key Performance Enhancements

Performance was boosted through:

- Advanced Data Augmentation: Employed **MixUp** (α=0.4) for robust regularisation and **geometric transforms** (random horizontal flip, crop, rotation) for data diversity.

- Optimised Training Strategies: Utilised **multi-stage progressive training** with **AdamW**, dynamic learning rate schedulers (OneCycleLR, ReduceLROnPlateau), and **label smoothing** for stable and effective learning.

2.3 Ensemble Strategy for Accuracy-Cost Trade-off

To achieve higher accuracy while maintaining efficiency, we employed **ensemble modelling**. We used **Weighted Average Probabilities (WAP)**, which consistently outperformed Reciprocal Rank Fusion. This approach allowed us to combine the strengths of multiple smaller models rather than relying on a single, larger, and more costly network. For our submission, we ensembled two distinct EfficientNet-B0 variants (one with geometric transforms, one with MixUp) with optimal weights **[0.45, 0.55]**.

3. Results and Performance Analysis

Our combined strategy yielded significant gains over the human baseline, showing a strong accuracy-cost balance.

3.1 Key Results

Model Configuration	Test Accuracy	Theoretical GFLOPs
Baseline Model	0.49675	0.32751
Average Human Performance	~0.6500	N/A
Best Single EfficientNet-B0 (Crop Rotate Flip)	**0.6666**	**0.0232**
EfficientNet-B0 (MixUp)	0.6468	0.0232
Submitted Ensemble (2-Model)	**0.6874**	**0.0464**
Highest Achieved Ensemble (5-Model)	0.6988	0.1161

Our **best single EfficientNet-B0 model** (*model_efficientnetb0_v2_tuned_2.pth*) achieved **0.6666 accuracy** at just **0.0232 GFLOPs**. The **submitted ensemble** (*model_efficientnetb0_v2_tuned_2.pth* and *model_efficientnetb0_v5_tuned.pth* with weights [0.45, 0.55]) reached **0.6874 accuracy** with a theoretical combined cost of **0.0464 GFLOPs**.

3.2 Discussion on Gains

These results demonstrate a substantial improvement over the baseline. Gains were attained by:

- **Leveraging EfficientNet-B0's inherent efficiency**, further enhanced by task-specific architectural adaptations that dramatically reduced FLOPs.
- **Robust data augmentation** strategies that boosted model generalisation on a challenging dataset.
- **Optimised training pipelines** that maximise individual model performance.
- **Strategic ensembling** of multiple efficient models, allowing us to combine diverse learned features for superior accuracy (0.6874) while maintaining a significantly lower computational profile (0.0464 GFLOPs) than a single, larger, and more complex model would require to achieve comparable performance. This validates our focus on the accuracy-cost trade-off.

4. Limitations and Conclusions

4.1 Limitations

While highly effective, our system did not achieve State-of-the-Art (SOTA) performance. Furthermore, to keep the competition fair, we only trained on the provided dataset and deliberately **did not train on external datasets like FER+**, despite their potential to significantly boost accuracy.

4.2 Conclusions

In conclusion, our project successfully delivered a robust and efficient solution for FER2013. By strategically adapting EfficientNet-B0, implementing advanced augmentation and training, and employing effective ensemble techniques, we achieved strong facial emotion recognition performance (0.6874 accuracy) within a practical computational budget (0.0464 GFLOPs). This demonstrates the viability of optimising smaller models and ensemble approaches for high performance without excessive resource demands.