Image Based Facial Emotion Recognition Using Deep Learning

Jaskirat Singh, Roxanne Mai

University of Calgary

Github: https://github.com/JaskiratSingh1/Image-Based-Facial-Emotion-Recognition-Using-Deep-Learning/tree/main

Abstract

Facial emotion recognition (FER) is a key area in computer vision and has applications in psychology, human-computer interaction, and security. In this project, we develop and evaluate a deep learning approach to classify human facial expressions into eight emotion classes: anger, content, disgust, fear, happy, neutral, sad, and surprise. We explored various models leveraging a publicly available dataset and a host of transfer learning pipelines with EfficientNetB2, MobileNetV2, or VGG16. Our best model with VGG16 achieves approximately 59% accuracy on a held-out test set. This report details our motivation, methodology, experimental results, and discussion of model performance, along with suggestions for further improvements.

Keywords – Convolutional neural network (CNN), deep CNN, emotion recognition, transfer learning

1. Introduction

Emotion recognition from facial images has gained attention in recent years for applications like mental health assessment, user experience personalization, and social robotics. Despite the success of deep learning in image classification, facial emotion recognition is challenging due to the subtlety of expressions, individual variability, and cultural influences.

Problem Statement: Our objective is to develop a model that classifies facial expressions into the correct emotion category given a single facial image. We utilize an existing dataset of labeled face images and explore the performance of a transfer learning approach.

Motivation and Significance: Accurately recognizing emotions can facilitate more natural human-computer interaction and enable new applications such as remote therapy or educational software that reacts to students' emotional states. As robots are increasingly utilized in social applications such as delivery services and autonomous vehicles, the accurate interpretation of human facial expressions is becoming increasingly critical. By improving the reliability of automated emotion detection, we can create more empathetic and effective technology-based solutions.

2. Related Work

Deep learning techniques – especially convolutional neural networks (CNNs) – have become the dominant approach for facial emotion recognition in recent years, surpassing earlier methods that relied on traditional handcrafted features[5]. For example, Ko's brief review in 2018 highlighted how CNN-based FER achieved state-of-the-art results with the availability of big training datasets [5]. More recently, comprehensive surveys (e.g. Li and Deng 2020 [7]) have documented the evolution of FER from controlled lab settings to in-the-wild scenarios, noting that advanced deep learning methods have rapidly emerged to tackle new challenges. These works emphasize that modern FER systems largely build on deep CNN architectures to automatically learn expressive features from face images, a strategy that has proven far more effective than traditional feature engineering.

Given the data-intensive nature of deep CNNs, transfer learning is commonly used in facial emotion recognition (FER) to address limited emotion-labeled data. Instead of training CNNs from scratch, many studies fine-tune pre-trained models (like ImageNet or VGG-Face) for emotion recognition [2]. This approach reuses learned features, improving accuracy and speeding up training for FER tasks [2]. For instance, Akhand et al. [2] adopted several pre-trained deep CNNs (including VGG-16, ResNet-50, Inception-v3, DenseNet-161, etc.), replacing their top layers and gradually fine-tuning each block on facial emotion datasets. Using this transfer-learning pipeline, they achieved state-of-the-art accuracies of 96.5% on KDEF and 99.5% on JAFFE with a DenseNet-161 model – a significant improvement over training a comparable network from scratch. Li et al. [6] showed that fine-tuning a pre-trained CNN can outperform custom architectures, even with small FER datasets, establishing transfer learning as a standard for static image FER. Our method aligns with this by initializing a deep CNN with pre-trained weights and retraining it on emotion data to leverage learned visual features. Recent work also explores model improvements alongside transfer learning. For example, some researchers integrate auxiliary features like facial action units or landmarks, or perform data augmentation to boost

performance [8]. Others have designed lighter CNNs tailored for FER: Debnath et al. [4] propose a compact four-layer "ConvNet" that still achieves about 96% training accuracy (and ~70% validation accuracy) on the FER2013 dataset by carefully optimizing training data diversity. This shows that, with the right strategy, even shallow networks can perform competitively on static FER tasks. Overall, the literature suggests that a well-chosen CNN architecture, enriched via transfer learning, is a highly effective solution for image-based emotion recognition.

Recent publications have further emphasized the effectiveness of CNN-based transfer learning for facial emotion recognition (FER). For instance, Agung et al. [1] proposed a dual approach in which CNN models were either trained from scratch or fine-tuned from pre-trained Inception-V3 and MobileNet-V2 architectures. Their system achieved up to 96% accuracy on a 10-emotion dataset. Behera and Shukla [3] extended the transfer learning paradigm by adopting multiple data sources to alleviate the few-shot challenge in FER. Their multi-source transfer learning approach maximizes correlated feature learning across different source datasets, thus reducing the risk of negative transfer. They reported an improvement of up to 7%–15% in accuracy compared to existing methods.

While most FER research focuses on static images, there is also a growing body of work dedicated to video-based emotion recognition. Liu et al. [9] introduced a Graph Convolutional Network (GCN) framework for video-based FER. By modeling temporal dynamics and relationships between facial landmarks across consecutive frames, they demonstrated improved accuracy over purely frame-by-frame CNN approaches. In a similar vein, Nie et al. [11] tackled multimodal emotion recognition by combining audio and video data via a correlation-based GCN, finding that fusing these complementary modalities can outperform vision-only pipelines.

Other recent efforts use standardized CNN architectures on static images but focus on making them more efficient for practical settings. Mehendale [10] presented a compact convolutional network for facial emotion classification, showing that even relatively shallow CNNs can achieve strong performance when properly configured. Meanwhile, Song et al. [12] emphasized real-time FER on smartphones by developing a deep learning pipeline optimized for on-device usage. Their solution illustrates how CNN-based methods can be deployed in resource-constrained environments without sacrificing robust facial expression detection.

These additional studies corroborate earlier findings that transfer learning, architectural optimizations, and careful dataset curation are key to achieving high accuracy and real-world viability in FER. Taken together, the literature strongly supports a transfer learning–based CNN strategy for static image classification, while also suggesting opportunities to expand into video or multimodal emotion recognition.

3. Materials and Methods

3.1. Dataset

We used a publicly available dataset (hosted on Hugging Face) containing 9,400 images labeled with 8 emotion classes: anger, content, disgust, fear, happy, neutral, sad, and surprise. Each image was accompanied by metadata indicating the emotion label. The dataset was divided into training, validation, and testing subsets (60% train, 20% validation, 20% test) using a stratified split to maintain class balance.

3.2. Data Preprocessing

1. Cleaning & Label Extraction: We parsed JSON-like metadata to extract the primary emotion label for each image. We removed any incomplete or corrupted records (none were found in this dataset).
2. Resizing & Normalization: Each image was resized to 224×224 pixels and scaled to [0, 1] by dividing pixel intensity values by 255.
3. Encoding Labels: We applied a LabelEncoder to transform categorical labels (e.g., "happy," "sad") into integer-encoded targets.
4. Data Augmentation: To improve the network's generalization and mitigate overfitting, we introduced a data augmentation strategy using TensorFlow's Keras API. The augmentation pipeline includes random horizontal flips, slight rotations (up to ±10°), random zooms, and random contrast adjustments. These augmentations simulate real-world variations in the dataset and enhance model robustness.

3.3. Model Architecture

We built our model as a Keras Sequential pipeline that integrates data augmentation into the network. The architecture was designed as follows:

- Input & Data Augmentation: The model accepts (224×224×3) images and includes an integrated data augmentation block that is active only during training.
- Feature Extraction: We explored three different pre-trained networks (all with ImageNet weights and include_top=False) as the base model:
    - EfficientNetB2 (Fine-tuned layers: 30 out of 340)

- MobileNetV2 (Fine-tuned layers: 30 out of 154)
- VGG16 (Fine-tuned layers: 10 out of 19)

- Classification Head:
  - A GlobalAveragePooling2D layer to reduce spatial dimensions.
  - A Dropout layer with a rate of 0.2 to mitigate overfitting.
  - A Dense layer (softmax) with 8 outputs, one for each emotion class.

We first froze the base model weights and trained only the new classification head. Afterward, we unfroze parts of the base model for fine-tuning with a lower learning rate to further improve performance.

3.4. Training Procedure

- Optimizer & Loss: We used the Adam optimizer with a default learning rate (1e-3) for the initial training and switched to a smaller learning rate (1e-5) during fine-tuning. The loss function was sparse_categorical_crossentropy.
- Hyperparameters: Each model (MobileNetV2, EfficientNetB2, VGG16) was trained for 50 epochs initially, followed by up to 35 epochs of fine-tuning, aiming to avoid overfitting through early stopping. The batch size was set to 32 throughout training.
- Early Stopping: We employed an early stopping mechanism that halts training when the validation loss does not improve for a specified number of epochs (patience). This was done to prevent overfitting and ensure that we retain the best model weights.
- Hardware: All experiments were conducted on a GPU-enabled environment (local GPU machine).

4. Results and Discussion

4.1. Quantitative Results

To evaluate the performance of our three transfer learning models — EfficientNetB2, MobileNetV2, and VGG16 — we tracked training and validation accuracy/loss curves and computed their test-set confusion matrices (Fig. 1–6). Each network was trained under the same pipeline (Section 3.4), including data augmentation and fine-tuning for 30–35 epochs on selected layers.

1. EfficientNetB2

- Accuracy & Loss Trends: The training graphs (Fig. 1) exhibit a slightly larger gap between training and validation accuracy compared to the other models, suggesting mild overfitting. However, careful fine-tuning of the deeper layers still led to improved overall performance by the final epochs.
- Confusion Matrix: (Fig. 2) Demonstrates moderate confusion among similar emotions (e.g., neutral vs. content, surprise vs. fear).

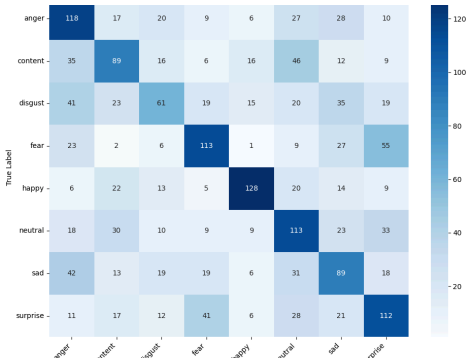

Fig. 1. Confusion Matrix for model with EfficientNetB2.

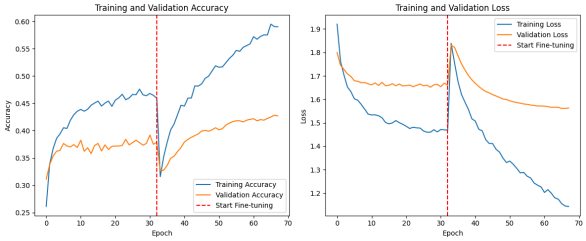

Fig. 2. Training and validation accuracy/loss curves of the model with EfficientNetB2.

2. MobileNetV2

- Accuracy & Loss Trends: The train-validation curves (Fig. 3) suggest slightly faster initial convergence than EfficientNetB2, possibly due to MobileNetV2's lighter architecture. However, the accuracy plateaued earlier, hinting that further hyperparameter tuning might be beneficial.
- Confusion Matrix: (Fig. 4) shows reduced confusion on neutral and content, but still struggles to cleanly separate surprise and sad.

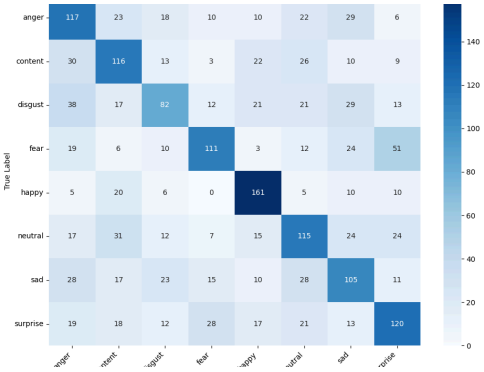

Fig. 3. Confusion Matrix for model with MobileNetV2.

Fig. 4. Training and validation accuracy/loss curves of model with MobileNetV2.

3. VGG16

- Accuracy & Loss Trends: From the training curves (Fig. 5), the model converged steadily with minimal overfitting, evidenced by a small gap between training and validation accuracy.
- Confusion Matrix: (Fig. 6) reveals the highest correct classification rates in general, while surprise and fear remain the most confused pair. The classification report shows that disgust is sometimes misinterpreted as anger, presumably due to shared facial muscle activation around the eyes.

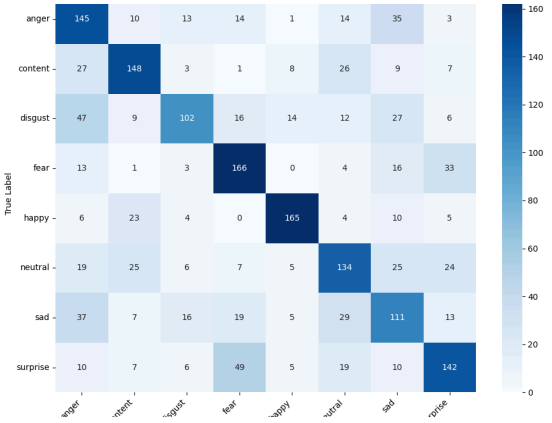

Fig. 5. Confusion Matrix for model with VGG16.

Fig.6. Training and validation accuracy/loss curves of the model with VGG16.

The table 1 below presents the final test loss and corresponding accuracy for each model. VGG16 achieves the highest test accuracy, possibly because its deeper architecture and feature extraction pipeline capture finer-grained emotional cues. MobileNetV2, although slightly behind in accuracy, shows faster convergence and a smaller memory footprint, indicating potential for real-time or mobile applications.

Table. 1. Models comparison with test loss and accuracy.

Model	Test Loss	Test Accuracy
EfficientNetB2	1.5070	43.03%
MobileNetV2	1.4313	49.31%
VGG16	1.1800	59.20%

4.2. Qualitative Analysis

A qualitative inspection of predictions provides insight into where each model succeeds or struggles. Below are some representative observations:

- Clear Expressions: All three networks reliably recognize emotions with strong, distinct facial cues—such as anger (characteristic lip curl) or happy (raised eyebrows, open mouth). Sample test images confirm that even with moderate lighting variations, these salient expressions are recognized with relatively high confidence.
- Subtle Emotions: Misclassifications are most frequent among neutral, disgust, and sad, which sometimes appear similar depending on the individual's facial structure or lighting conditions. For instance, multiple test samples labeled as neutral were predicted as disgust by the MobileNetV2 (Fig. 7). VGG16 also occasionally confuses anger with sad, suggesting certain overlapping features (e.g., tightened lips, furrowed brows) (Fig. 8).
- Impact of Augmentation: From the training data, we see that random flips and zooms help the models generalize better to rotated or slightly off-center faces.
- Model Confidence: When classifying images with prominent emotional features, all three CNNs display high confidence scores. In borderline cases with partial facial visibility (side profile or partial occlusion), the confidence drops and misclassification rates rise, consistent with prior studies on occlusion handling (Fig. 9).

Fig. 7. Five samples of qualitative inspection of predictions from the model with MobileNetV2.

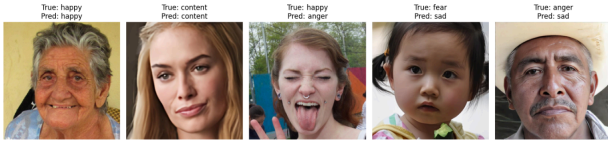

Fig. 8. Five samples of qualitative inspection of predictions from the model with VGG16.

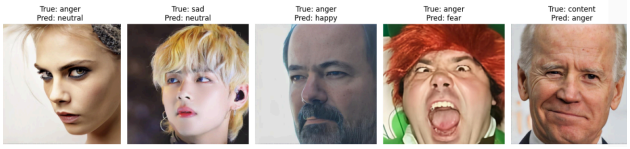

Fig. 9. Five samples of qualitative inspection of predictions from the model with VGG16.

4.3. Limitations and Strengths

Strengths:

- The dataset is fairly diverse, covering 8 different emotions. It helps the model generalize better.
- Data preprocessing techniques like augmentation make the model more robust and reduce overfitting.
- Using multiple pre-trained models (EfficientNetB2, MobileNetV2, VGG16) allows for easy performance comparisons and adaptability to different uses.
- An early stopping mechanism helps prevent overfitting and keeps the best model weights.
- Clear evaluation metrics, such as confusion matrix, test loss and accuracy, make it simple to assess performance and compare models.

Limitations:

- The dataset may not include enough cultural and demographic diversity, which can lead to biases in emotion recognition.
- The models struggle to accurately classify subtle emotional expressions, especially when emotions are similar (e.g., neutral vs. disgust).
- Models require significant computational resources, which can limit their use in real-time applications.
- Deeper models, such as EfficientNetB2, may overfit if not managed properly, indicated by gaps between training and validation accuracy.
- Confusion among similar emotions shows a need for better feature extraction and data labeling to improve model differentiation.

5. Conclusions

In this work, we presented a transfer learning–based pipeline for recognizing facial emotions from images across eight categories: anger, content, disgust, fear, happy, neutral, sad, and surprise. Our approach builds on established deep convolutional neural networks – specifically EfficientNetB2, MobileNetV2, and VGG16 – each pre-trained on ImageNet and fine-tuned for emotion classification. To reduce overfitting, we employed data augmentation (random flips, rotations, zooms, and contrast adjustments) and applied early stopping criteria to preserve optimal weights.

From the experiments summarized in, VGG16 exhibited the highest test accuracy at approximately 59%, surpassing both MobileNetV2 and EfficientNetB2 within a similar number of epochs (Table. 1). This result affirms findings in the literature that deeper architectures, when carefully fine-tuned, can better capture the subtle features distinguishing closely related emotions. Overall performance was hindered by the complexity of the dataset, potential class imbalances, and nuances between some expressions such as content vs. neutral.

Despite these challenges, our pipeline demonstrates the feasibility of efficiently recognizing a wide range of emotional states using transfer learning. The results suggest that future work could focus on:

- Hyperparameter Tuning: Systematically refining training parameters such as batch size, optimizer type, or learning rate schedules to improve performance.
- Advanced Data Augmentation: Introducing occlusion-based or pose variation augmentations for greater robustness to real-world variability.
- Ensemble Methods: Combining the outputs of multiple CNN models (e.g., MobileNetV2 + VGG16) to capture complementary strengths.
- Temporal & Multimodal Fusion: Exploring video-based FER or fusing other modalities like audio for more comprehensive emotion understanding, in line with state-of-the-art approaches.

Overall, the proposed system – anchored by transfer learning on CNN backbones – provides a strong foundation for facial emotion recognition, demonstrating consistent gains over training from scratch and highlighting the potential for further enhancement with complementary techniques and additional data.

6. References

[1] G. C. Agung, L. T. Bariyah, A. S. R. Dwifantara, and E. S. Iswahyudi, "Detecting Emotions from Facial Images Using Transfer Learning-Based CNN Approaches," *Sensors*, vol. 23, no. 20, Art. no. 8641, Oct. 2023, doi: 10.3390/s23208641.

[2] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, vol. 10, no. 9, Art. no. 1036, Apr. 2021, doi: 10.3390/electronics10091036.

[3] S. Behera and A. Shukla, "Multi-Source Transfer Learning for Few-Shot Facial Emotion Recognition," *Sensors*, vol. 23, no. 17, Art. no. 7404, 2023, doi: 10.3390/s23177404.

[4] M. Debnath, M. T. H. B. Ahmed, M. S. Islam, M. Rahman, and M. F. Ali, "A Real-Time Facial Emotion Recognition System Using a Lightweight Convolutional Neural Network," *Sensors*, vol. 22, no. 9, Art. no. 3429, Apr. 2022, doi: 10.3390/s22093429.

[5] B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors*, vol. 18, no. 2, Art. no. 401, Jan. 2018, doi: 10.3390/s18020401.

[6] T. Y. Li, H. X. Li, S. J. Wei, and F. Chen, "An Anti-Aliased Deep Convolution Network Model for Robust Facial Emotion Recognition," *Sensors*, vol. 23, no. 7, Art. no. 3602, Apr. 2023, doi: 10.3390/s23073602.

[7] X. Li and S. Deng, "Deep Facial Expression Recognition: A Survey," *arXiv Preprint*, arXiv:1804.08348, 2018. [Online]. Available: http://arxiv.org/abs/1804.08348

[8] C. F. Liew and T. Yairi, "Facial Expression Recognition and Analysis: A Comparison Study of Feature Descriptors," *IPSJ Trans. Comput. Vis. Appl.*, vol. 7, pp. 104–120, 2015. [Online]. Available: https://www.jstage.jst.go.jp/article/ipsjtcva/7/0/7_104/_article/-char/ja/

[9] D. Liu, H. Zhang, and P. Zhou, "Video-based Facial Expression Recognition Using Graph Convolutional Networks," in *Proc. 2020 25th Int. Conf. Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 607–614, doi: 10.1109/ICPR48806.2021.9413094.

[10] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Appl. Sci.*, vol. 2, Art. no. 446, Feb. 2020, doi: 10.1007/s42452-020-2234-1.

[11] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation Based Graph Convolutional Network for Audio-Video Emotion Recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3793–3804, 2021, doi: 10.1109/TMM.2020.3032037.

[12] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *Proc. 2014 IEEE Int. Conf. Consumer Electronics (ICCE)*, Las Vegas, NV, USA, 2014, pp. 564–567, doi: 10.1109/ICCE.2014.6776135.