FACIAL EMOTION RECOGNITION USING MOBILENET BASED TRANSFER LEARNING

Subhashree Kedia Samriddhi Varshney Apoorva Neha

***Abstract*—This study explores the use of MobileNet for facial emotion recognition (FER) with a focus on addressing class imbalances and fine-tuning the model's classifier. To counteract dataset imbalances, we performed data augmentation and removed underrepresented emotions. The baseline MobileNet model achieved 49 percent accuracy, while a fine-tuned version with additional classifier layers reached 48 percent, suggesting that adding layers did not improve performance. A basic CNN model was also tested, reaching 51 percent accuracy. Despite these moderate results, MobileNet shows promise for real-time FER applications, including mental health monitoring, accessibility tools, and driver safety. This research demonstrates MobileNet's strengths and limitations in FER, highlighting avenues for future improvements in model optimization**

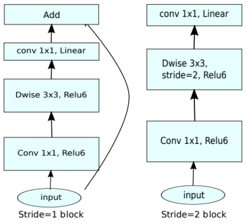
***Index Terms*—CNN, Facial Emotional Recognition, Pre-trained MOBILENET.**

I. Introduction

1. Facial Emotion Recognition (FER) involves detecting and analyzing human emotions through facial expressions. With applications across human-computer interaction, social robotics, mental health monitoring, and entertainment, FER technology holds significant potential for advancing digital communication and user experience. However, deploying FER on mobile devices poses challenges due to the limited computational power, storage, and battery life in these devices. This project aims to develop an FER system that is both accurate and efficient enough for real-time performance on mobile devices.

2. Facial Emotion Recognition (FER): FER systems are designed to recognize basic emotions (e.g., happiness, sadness, anger) by analyzing facial images. Traditional FER methods rely on feature extraction techniques like Gabor filters and histogram of oriented gradients (HOG) combined with machine learning classifiers. With the rise of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach, excelling in image recognition tasks by learning hierarchical features directly from raw data.

3. Introduction to MobileNet: MobileNet, an efficient and compact CNN architecture. By employing depthwise separable convolutions, MobileNet dramatically reduces the number of parameters and computations, making it faster and more suitable for mobile devices without sacrificing significant accuracy. MobileNet's hyperparameters, like the width and resolution multipliers, allow fine-tuning of the model to achieve a suitable trade-off between latency and performance. 4. Research Objectives: This research focuses on creating a real-time FER model that is both lightweight and effective on mobile devices. The key objectives include: - Leveraging MobileNet for real-time emotion classification. - Employing transfer learning to fine-tune MobileNet for FER. - Balancing efficiency and accuracy to meet the constraints of mobile devices.



(d) Mobilenet V2

5. Motivation: The use of MobileNet in this project is motivated by its balance between size, speed, and performance. Efficient deployment of FER on mobile devices enables real-time analysis, making FER accessible in a variety of settings without dependence on cloud resources. Real-time FER on mobile could benefit applications where user privacy, latency, or connectivity constraints are critical, such as in remote health monitoring or interactive mobile apps.

II. Literature Review

1. Previous Research on FER : Early FER techniques were based on image feature extraction, often using traditional computer vision methods like Gabor filters and HOG. These methods required manual feature engineering, which was computationally intensive and less effective on varied datasets. The introduction of CNNs transformed FER, enabling automatic feature extraction and achieving high accuracy by learning hierarchical representations from large datasets. CNNs outperform traditional methods, particularly in cases where expressions vary due to lighting, occlusion, or facial diversity.

2. Challenges in FER: Key challenges in FER include: - **Cultural Differences**: Facial expressions vary across cultures, which can lead to misinterpretation if the dataset is not diverse. - **Real-Time Constraints**: FER must meet high processing speeds, especially on mobile devices, for applications like interactive user feedback. - **Data Limitations**: FER models require extensive labeled data for effective training, but gathering a large and diverse dataset is challenging. Privacy concerns also arise when collecting facial data.

3. Comparative Studies: MobileNet's performance has been evaluated against other models employing CNN. MobileNet, with fewer parameters and optimized design, achieves a favorable trade-off, especially suitable for mobile devices where latency and efficiency are paramount. Some studies have even integrated MobileNet with additional modules, such as attention layers, to further enhance FER accuracy.

III. Methodology

* Choice of Dataset- the dataset was chosen from Kaggle the dataset had 8 emotions categories – anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise. All images in it contain grayscale human face (or sketch). Each image was 224 x 224 pixel grayscale in PNG format.

1. Preprocessing Steps In this code, the preprocessing pipeline involves preparing batches of images for a neural network (mobilenet) to perform emotion classification. The key preprocessing steps are as follows:

• Batch Processing: Images are processed in batches of 16 (batch size = 16) to optimize GPU/CPU usage and enable efficient inference. A batch approach helps reduce memory usage and speeds up the processing.

• Image Loading: Each image is loaded from its file path from the Pillow library.

• RGB Conversion: All images are converted to RGB format to ensure consistent color channels, which is required for the mobilenet model, as it expects three-channel (RGB) images.

• Resizing: Each image is resized to the model's expected input dimensions, typically 224x224 pixels for MobileNet.

• Normalization: Pixel values are normalized to fit the model's expected input distribution. For MobileNet, this typically involves scaling values to a standard range (e.g., between 0 and 1) and normalizing each channel with specific mean and standard deviation values.

• Tensor Conversion: Images are converted into PyTorch tensors, and the batch of tensors is then combined into a single tensor with a shape [batchsize, 3, 224, 224]for batch processing in the model.

•Label Mapping: Predicted class indices are mapped to emotion labels (e.g., ”happy”, ”sad”) using a dictionary created from the label encoder's classes. If a predicted class index is not found in the mapping, a warning is issued, and ”unknown” is added to the predictions.

2. Data Augmentation:

• Class Imbalance: In the dataset a class imbalance was noticed among the emotions. Some of the emotions like contempt had very less data compared to others.

• Data Augmentation and Prepration: Using this we try to improve the class balance by creating new samples from the existing data, increasing the size of the class. this also introduces additional variations in the data, which can help prevent overfitting by providing the model with a more diverse set of samples.

The samples were increased to a size of 800 then 900 and finally 1000 which was found optimal for model training.

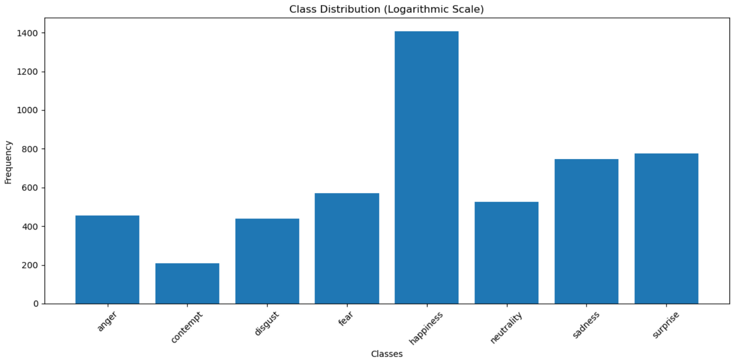


Fig. 1: Imbalance in data

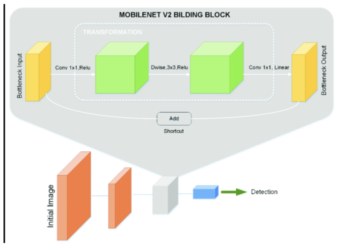


Fig. 2: MOBILENET V2

• Removing classes: Some classes like contempt with very less data were deleted as their augmentation contributed to bad accuracy.

• At last, only 5 major classes were included for better accuracy.

3. Mobilenet Architecture: The architecture centers on depthwise separable convolutions and inverted residual blocks with linear bottlenecks, which reduce parameters while preserving detailed features. Depthwise convolutions filter each channel independently, followed by pointwise convolutions to combine channel information. Inverted residual blocks expand low-dimensional inputs to a higher dimension before projecting them back down, preserving information while maintaining efficiency. The linear bottleneck layer prevents data loss by avoiding non-linear activations post-projection. MobileNetV2 concludes with global average pooling and a fully connected layer, which can be fine-tuned for specific tasks.

4. Model Training: For training the facial emotion recognition model, we utilized the **MobileNetV2** architecture, a pre-trained convolutional neural network known for its efficiency and high accuracy in lightweight applications. To adapt MobileNetV2 for our specific task, we modified the final classification layer to match the number of emotion categories, mapped through a label encoding process. By freezing the initial layers of MobileNetV2 and selectively unfreezing the classifier layers, we fine-tuned the model to recognize emotions with enhanced accuracy while maintaining efficiency. The training was conducted using **Adam optimization** with a batch-wise gradient descent approach.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \cdot g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) \cdot g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

where g_t is the gradient at step t, θ represents model parameters, and η is the learning rate. We applied Cross-Entropy Loss as the criterion for updating weights, leveraging batch-based processing to improve computational efficiency. The cross-entropy loss is defined as

$$\mathscr{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\hat{y}_{ij})$$

, where N is the number of samples, C is the number of classes, y_{ij} is the true label, and $\hat{y}_{ij}$ is the predicted probability.

Training spanned 10 epochs, during which key performance metrics like loss and accuracy were logged to monitor model progress. describe the deployment of the model on mobile hardware.

4. Parallelly training a CNN Model:

We trained a CNN model on tha same augmented data to compare the speed and accuracy with that using mobilenet.

• Model Architecture:

We constructed a Convolutional Neural Network (CNN) named EmotionCNN. The architecture consists of two convolutional layers followed by max-pooling layers. The first convolutional layer uses 32 filters, while the second uses 64 filters. After passing through these layers, the feature maps are flattened and passed through two fully connected layers. The output layer produces predictions for five emotion categories, aligning with our classification task.

• Training Process:

The model was trained using the Adam optimizer, which is well-suited for this type of task, and the Cross-Entropy loss function, which is commonly used for multi-class classification problems. Over the course of ten epochs, the training loop processes batches of images, calculates the loss, and updates the model parameters through backpropagation. At the end of each epoch, the average loss is printed to monitor the training progress.

5. Accuracy:

Here's a more concise version of your text, maintaining the essential points while reducing verbosity:

The accuracyscore function computes the overall accuracy of the model by comparing predicted labels to true labels across test samples. Accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

This metric reflects the proportion of correctly classified facial emotions, but may be misleading in cases of class imbalance.

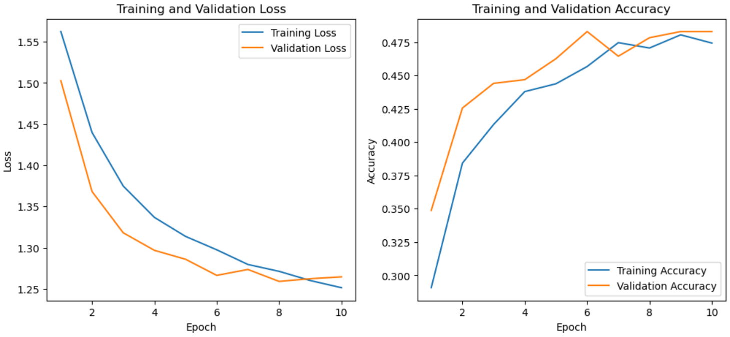


Fig. 3: Accuracy and Loss plots

The classificationreport offers a detailed analysis of model performance for each emotion class, calculating key metrics:

Precision indicates the accuracy of predicted instances:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Recall measures the model's ability to identify actual instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

F1-Score is the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

These metrics are computed for each class, with overall "macro" or "weighted" averages summarizing performance across all emotion categories

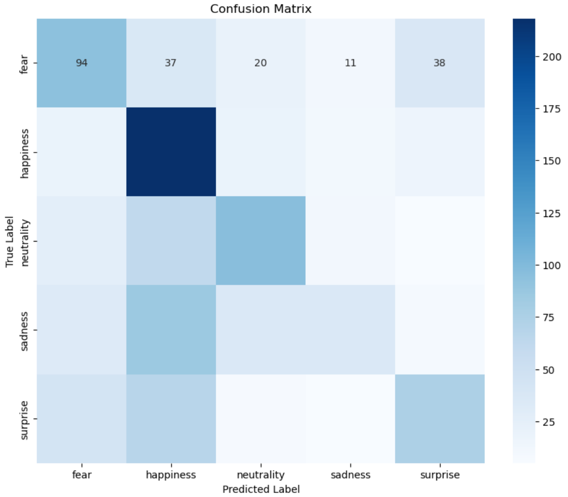
IV. Results and Discussions

A. Results

In our evaluation of model performance, we found that the fine-tuned MobileNetV2 achieved an accuracy of 49

The training of MobileNetV2 required approximately 34 minutes, significantly less than the CNN, which took around 60 minutes to train. This disparity in computational time indicates that MobileNetV2 is not only efficient in terms of its architecture but also optimized for speed, making it suitable for applications where quick training is essential.

Considering that both models yield similar accuracy levels, the choice of MobileNetV2 becomes more favorable due to its faster training time. This efficiency can be particularly advantageous in real-time applications, such as mobile and embedded systems, where computational resources are often limited. Thus, MobileNetV2 demonstrates a compelling balance between performance and computational efficiency, making it a strong candidate for practical implementations in facial emotion recognition tasks.



B. Discussions

This study investigated the use of MobileNet for Facial Emotion Recognition (FER), addressing class imbalance in the dataset and experimenting with model fine-tuning to optimize performance. The initial dataset exhibited significant class imbalances, with certain emotions underrepresented. To tackle this, we removed certain emotions and applied targeted data augmentation to balance the dataset.

We first evaluated the baseline MobileNet model without any modifications, which achieved an accuracy of 49%. To explore potential performance improvements, we then fine-tuned MobileNet by adding extra layers to the classifier, specifically designed for our emotion classes. However, this modified model produced a slightly lower accuracy of 48%, suggesting that adding layers did not improve FER performance in this context. We also benchmarked a basic CNN model, which achieved a comparable accuracy of 51%.

The trends observed in accuracy and loss across the 10 epochs, with data split into 60% training, 20% validation, and 20% test sets, indicated moderate model convergence but highlighted the limitations of the MobileNet architecture for this particular FER dataset. These findings suggest that while MobileNet's lightweight architecture is suitable for real-time applications, further optimization may be necessary to handle nuanced emotion classification effectively.

Potential applications for a MobileNet-based FER system remain extensive, even with these moderate accuracy results. For instance, mental health applications could use it to monitor mood changes over time, while accessibility tools might support individuals with autism in recognizing social cues. Additionally, driver safety systems could employ this technology to detect distraction or drowsiness, enhancing real-time monitoring in high-stakes environments.

V. Conclusions and Future Scopes

A. Conclusion

This research examined the feasibility of using MobileNet for facial emotion recognition. Despite attempts to improve performance by adding layers to the classifier, the fine-tuned model achieved a slightly lower accuracy (48%) compared to the baseline MobileNet (49%). This outcome highlights MobileNet's effectiveness as a real-time FER tool in its original form but suggests that the added complexity of extra layers did not contribute positively in this case. The comparable accuracy of a basic CNN model (51%) further emphasizes the challenges of achieving high precision in FER.

MobileNet's efficiency and adaptability for real-time FER applications make it valuable for diverse scenarios, from driver monitoring to accessibility tools. The model's moderate accuracy suggests that future work could focus on refining model architecture or exploring alternative methods to improve emotion classification in real-world settings.

B. Future Scopes

1. Dataset Expansion: Future research can benefit from the expansion of FER datasets that cover more varied facial expressions across different cultural and demographic groups. Improved datasets would help reduce cultural bias and enhance the generalizability of FER models. Including more subtle emotional expressions can also improve model robustness in real-world scenarios.

2. Optimization for Edge Devices: Although MobileNet is optimized for mobile, further improvements can enhance its deployment on edge devices with limited resources. Techniques like model quantization, pruning, and edge-specific optimization could be applied to reduce model size and power consumption. Efficient implementation of MobileNet with hardware accelerators like Tensor Processing Units (TPUs) on mobile could also be explored to further reduce latency.

3. Multimodal Emotion Recognition: FER systems could be integrated with other data sources, such as voice or physiological data, to create multimodal emotion recognition systems. Combining visual data with audio or physiological signals could provide a more holistic view of user emotions, especially in cases where facial expressions alone may not fully capture the emotional state. This approach could enhance emotion recognition, particularly in complex emotional scenarios.

4. Adaptive Learning Techniques: For FER systems deployed on mobile devices, adaptive learning techniques could enable continuous learning from new data. By updating the model periodically based on new inputs, FER models could maintain accuracy in dynamic environments without requiring retraining from scratch. Techniques such as federated learning could also allow for distributed training while maintaining data privacy, which is crucial for applications in healthcare or education.

References

[1] EPRA International Journal of Research and Development (IJRD) Volume: 8,Issue: 7,July 2023 - Peer Reviewed Journal on Facial Emotional Recognition by Jenisha A, Aleesha Livingston [2] https://arxiv.org/abs/1704.04861