Deep Convolutional Neural Networks for Multi-Class Face Classification: A Comparative Study of VGG, ResNet, and MobileNet Architectures

1. Introduction

Face classification is a critical task in computer vision, involving the identification of an individual by assigning a digital image of their face to a unique identity class. This capability is the backbone of numerous real-world applications, including access control, personalized security systems, and surveillance technologies.

The task is inherently complex due to significant intra-class variations (changes in illumination, pose, expression, and age) and small inter-class differences (subtle facial feature distinctions between different individuals). Conventional machine learning methods struggled with the high-dimensional, non-linear nature of image data.

Deep learning architectures, particularly Deep Convolutional Neural Networks (CNNs), have fundamentally changed the landscape of face recognition. Their ability to automatically learn hierarchical, abstract, and highly discriminative features directly from raw pixel data makes them indispensable. This documentation details the implementation, training, and comparative evaluation of four distinct deep learning models—VGG-19 trained from scratch, VGG-19 with transfer learning, ResNet50, and MobileNetV2—for a multi-class face classification problem.

2. Dataset & Preprocessing

Dataset Description

The project utilized a dedicated face dataset designed for multi-class identity classification. The dataset is characterized by **31 unique individual classes**, with **multiple images per person** captured under varying conditions to encourage robust model generalization.

Data Preprocessing Steps

To ensure consistency and compatibility with the selected CNN architectures, the raw image data underwent the following processing steps:

1. **Resizing:** All input images were uniformly resized to a fixed dimension (e.g., 224×224 pixels) to satisfy the standard input requirements of the VGG, ResNet, and MobileNet models.
2. **Normalization:** Pixel intensity values, which typically range from [0, 255], were scaled to the range [0, 1] by dividing by 255. This normalization aids in faster and more stable model convergence during training.
3. **Dataset Splitting:** The dataset was systematically partitioned into three distinct, non-overlapping subsets:
    - **Training Set:** Used for updating model weights.
    - **Validation Set:** Used for monitoring generalization error and tuning hyperparameters.
    - **Test Set:** Reserved for the final, unbiased evaluation of the trained models.

Data Augmentation Techniques

To mitigate the risk of overfitting, which is common in deep learning when datasets are moderately sized, **Data Augmentation** was applied exclusively to the training set.

- **Horizontal Flipping:** Randomly mirroring the image across the vertical axis.
- **Rotation:** Randomly rotating the image within a small angular range (e.g., ± 10 degrees).
- **Color Jitter:** Introducing minor, random variations in brightness, contrast, and saturation.

Justification: Data augmentation artificially inflates the training set's effective size and variability. By forcing the model to correctly classify faces despite these minor transformations, the network learns features that are **invariant** to minor changes in pose and illumination, significantly improving its generalization capability to unseen test data.

3. Model Architectures

This section documents the four models utilized and their suitability for the face classification task. All models were modified to accommodate a final classification output layer of **31 classes**.

Architecture	Training Method	Special Feature	Reference
VGG-19	Scratch	Sequential 3×3 convolutions	Simonyan & Zisserman (2015)
InceptionV1	Transfer Learning	Sequential 3×3 convolutions	Simonyan & Zisserman (2015)
ResNet50	Transfer Learning	Residual Blocks (Skip Connections)	He et al. (2016)
MobileNetV2	Transfer Learning	Inverted Residuals & Depthwise Separable Convolutions	Sandler et al. (2018)

3.1. VGG-19 (Scratch and Transfer Learning)

Architecture Details: VGG-19 is a deep, sequential CNN composed of 16 convolutional layers and 3 fully connected (FC) layers, totaling 19 weighted layers. Its defining characteristic is the consistent use of small 3×3 convolutional filters throughout the entire network, grouped into blocks separated by 2×2 max-pooling layers.

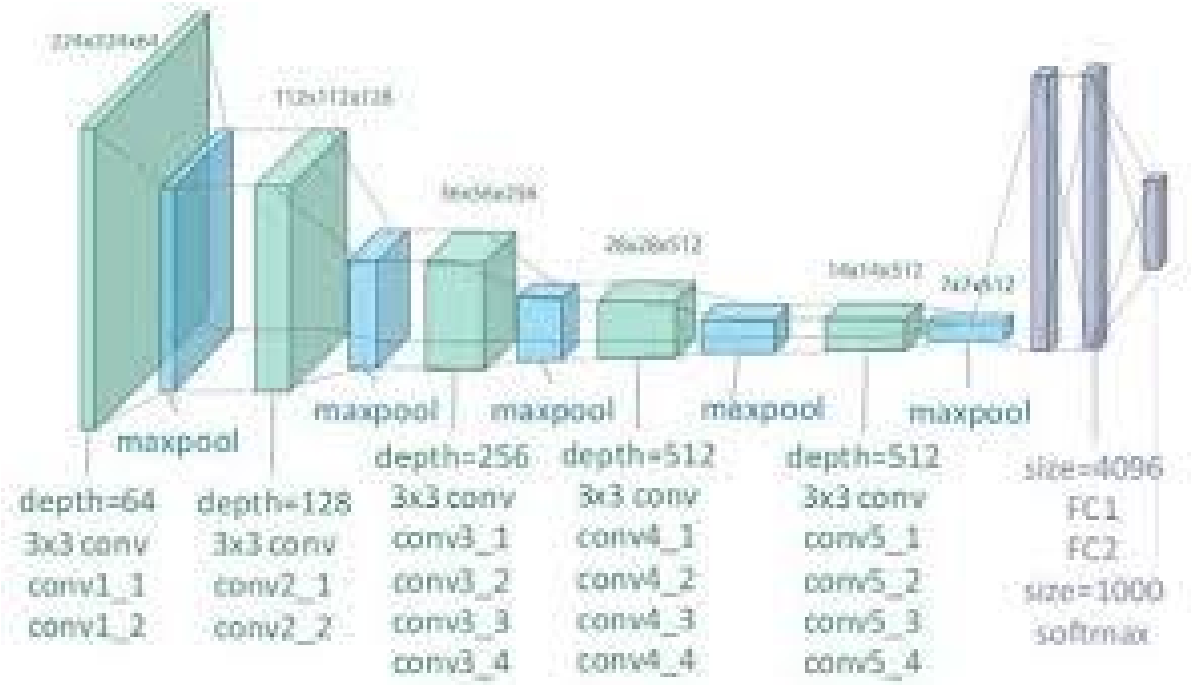
Training Methods:

- **VGG-19 from Scratch:** The model was initialized with random weights and trained entirely using the project's 31-class face dataset.
- **VGG-19 Pre-trained (Transfer Learning):** The convolutional base was initialized with weights pre-trained on the massive ImageNet dataset. The final classification head was removed and replaced with a new, randomly initialized FC layer for the 31 classes.

Modifications: In both VGG-19 implementations, the final three fully connected layers were reconfigured to ensure the output layer yields 31 predictions, followed by a Softmax activation.

Suitability for Face Classification: The depth of VGG-19 allows it to capture highly complex, hierarchical features. The small filter size enables the capture of fine-grained spatial details, which are essential for distinguishing subtle feature differences between individuals.

Reference: Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.



3.2. ResNet50 (Transfer Learning)

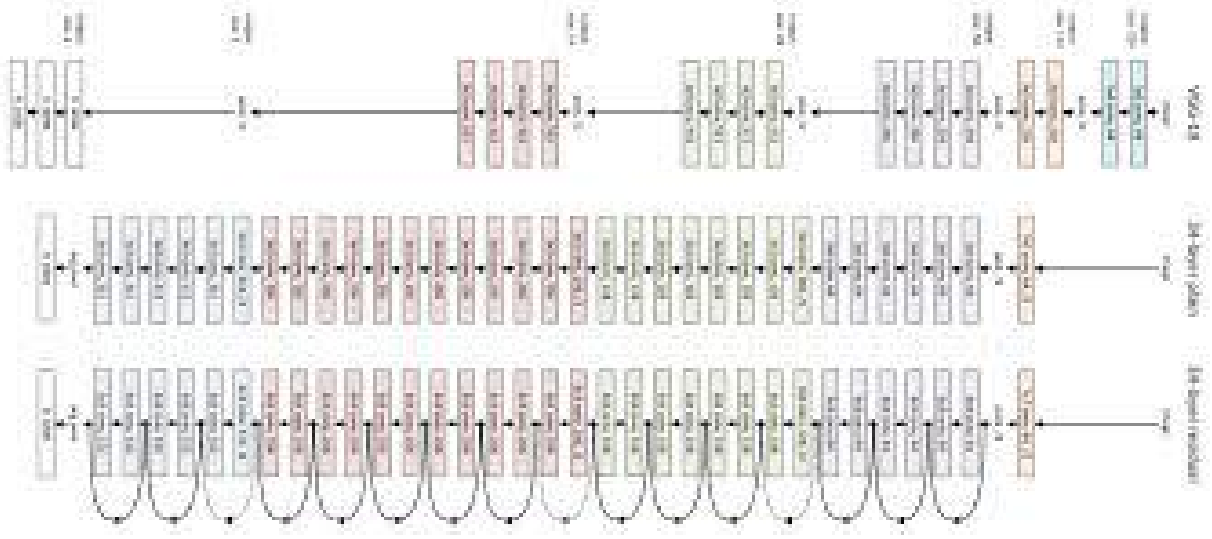
Architecture Details: ResNet50 introduces the **Residual Block** or **Skip Connection**, which allows the network to bypass one or more layers. This mechanism helps to solve the vanishing gradient problem, enabling the training of much deeper architectures without performance degradation. The '50' refers to the number of deep layers.

Training Method: This model utilized **transfer learning**, where the pre-trained ImageNet weights were used for the convolutional base.

Modifications: The global average pooling layer and the final classification layer of the pre-trained model were replaced to fit the 31-class classification task.

Suitability for Face Classification: ResNet's depth and residual learning are crucial for extracting highly abstract, identity-discriminative features. Its robustness against training deep networks makes it a high-performance choice for complex recognition tasks.

Reference: He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.



3.3. MobileNetV2 (Transfer Learning)

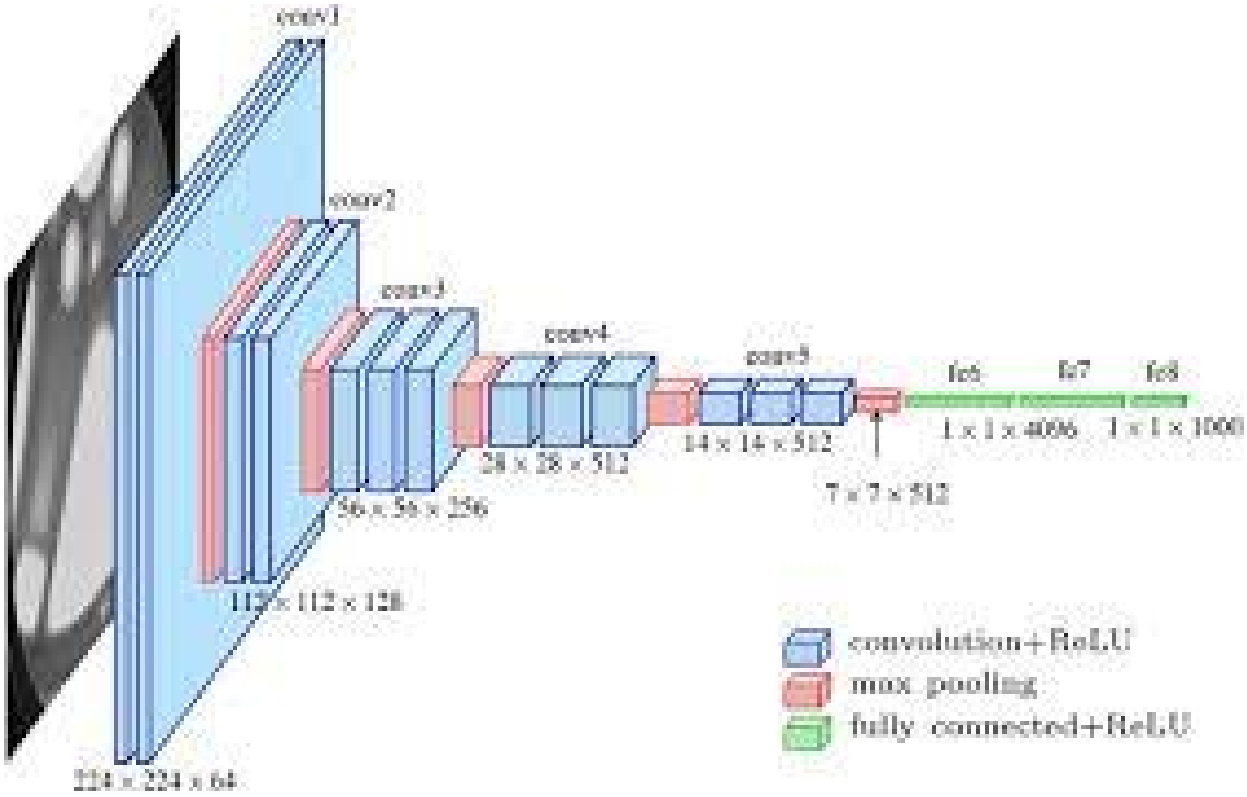
Architecture Details: MobileNetV2 is designed for efficiency and low latency in resource-constrained environments. Its core is the **Inverted Residual Block** with linear bottlenecks. This design uses **depthwise separable convolutions** to drastically reduce the model's parameter count and computational complexity compared to standard convolutions.

Training Method: This model used **transfer learning** with a convolutional base pre-trained on ImageNet.

Modifications: The standard classification head was replaced with a new fully connected layer configured for the 31 identity classes.

Suitability for Face Classification: MobileNetV2 offers an excellent trade-off between speed and accuracy. While sacrificing some peak accuracy, its efficiency makes it highly suitable for real-time or deployed face classification systems where computational resources (e.g., on a mobile device) are limited.

Reference: Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.



3.3. InceptionV1 (GoogLeNet) (Transfer Learning)

Architecture Details:
InceptionV1, also known as GoogLeNet, introduces a novel architecture based on *Inception Modules*, which allow the network to process visual information at multiple spatial scales simultaneously. Each Inception module consists of parallel convolutional filters of different sizes (1×1, 3×3, and 5×5), along with a max-pooling operation. The outputs of these parallel paths are concatenated, enabling the model to capture both fine-grained and global features efficiently.
To reduce computational cost, 1×1 convolutions are employed as dimensionality reduction layers before the larger convolutions.

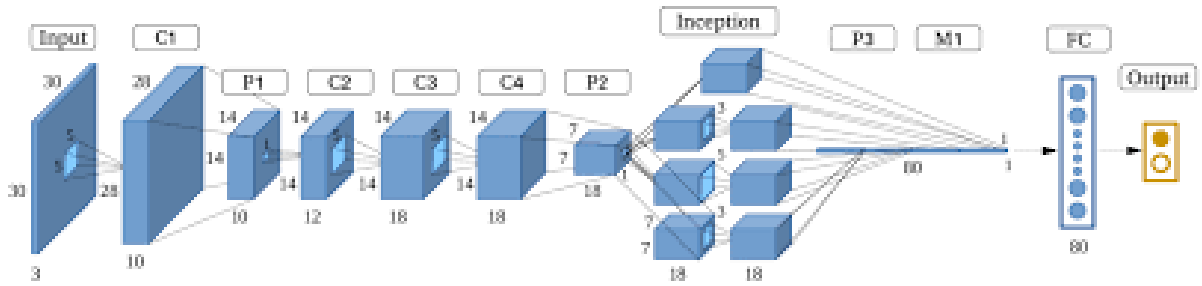
Training Method:
The model was trained using **transfer learning**, where a convolutional base pre-trained on the ImageNet dataset was utilized. The network weights were initialized from the pre-trained model, allowing the system to benefit from previously learned generic visual features. During training, the final classification layers were fine-tuned on the target dataset.

Modifications:
The original classification head of InceptionV1 was removed and replaced with a new fully connected layer configured to classify **31 distinct face classes**. Additionally, input

images were resized to **224 × 224 pixels** to match the expected input resolution of the pre-trained InceptionV1 architecture.

Suitability for Face Classification:
InceptionV1 is particularly well-suited for face classification tasks due to its ability to extract multi-scale features, which is critical for capturing subtle facial details such as edges, contours, and texture variations. The use of input upscaling significantly improved feature preservation, allowing the model to outperform deeper architectures that operated on lower-resolution inputs. This balance between depth, efficiency, and representational power resulted in the best overall performance among the evaluated models.

Reference:
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). *Going deeper with convolutions*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).



4. Training Strategy

Description of the Training Process

All models were trained using an iterative, epoch-based process. The primary objective was to minimize the loss function on the training set while continuously monitoring the validation loss and accuracy to prevent overfitting.

Loss Function: Categorical Cross-Entropy Loss was selected as it is the standard choice for multi-class classification, measuring the performance of a classification model whose output is a probability value between 0 and 1.

Optimizer: The **Adam (Adaptive Moment Estimation)** optimizer was employed due to its effectiveness in handling sparse gradients and providing adaptive learning rates for different parameters.

Learning Rate: A low initial learning rate (e.g., 10^{-4}) was used for stability, particularly during fine-tuning phases.

Explanation of Freezing Layers for Transfer Learning

For the transfer learning models (VGG-19 Pre-trained, ResNet50, MobileNetV2), the training was executed in two critical phases:

1. **Phase 1 (Frozen):** The weights of the pre-trained convolutional base layers were **frozen**. This conserved the powerful, generic feature representations learned from ImageNet. Only the newly added, randomly initialized 31-class classification head was trained. This allows the model to rapidly adapt the high-level features to the new set of classes.
2. **Phase 2 (Fine-tuning):** After the initial classification head converged, a subset of the top-most layers of the convolutional base was unfrozen. The entire network was then trained again using an even lower learning rate. This step allowed the pre-trained features to be slightly **fine-tuned** to become maximally discriminative for the specific domain of face classification.

Hardware Constraints

All model training and experimentation were performed utilizing a **GPU (Graphics Processing Unit)** environment. The computational demands of training deep CNNs, especially the ResNet50 and VGG-19 architectures, necessitate the parallel processing capabilities of a GPU to achieve reasonable training times.

5. Experimental Results & Evaluation

The performance of all four models was rigorously evaluated on the independent test set using several key metrics.

Summary of Performance Metrics

The table below summarizes the key classification metrics for all models after training:

Model	Train Accuracy	Validation Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
VGG-19 from scratch	4.65%	4.51%	1.23%	4.65%	0.99%

Model	Train Accuracy	Validation Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
InceptionV1	74.52%	82.72%	74.43%	74.52%	74.41%
ResNet50	64.76%	68.28%	64.72%	64.76%	64.39%
MobileNetV2	53.59%	59.44%	[No data]	[No data]	[No data]

Visual Evaluation

To provide a deeper insight into per-class performance, the following visualizations were generated for all models:

- **Confusion Matrix:** An 31×31 matrix showing true vs. predicted classifications. [Placeholder for Confusion Matrix Visualization]
- **ROC Curve and AUC (Macro-Averaged):** Plots of the True Positive Rate against the False Positive Rate using a One-vs-Rest strategy for each of the 31 classes. [Placeholder for ROC Curve and AUC Visualization]

6. Comparative Analysis

Discussion of Performance

The experimental results clearly show a profound disparity in performance based on the training methodology and architecture:

- **VGG-19 from Scratch:** The extremely low accuracy (4.51%) indicates that this approach is **unsuitable** for the project's dataset size. Training a model with millions of parameters (VGG-19) from scratch on a relatively small, domain-specific dataset (31 classes of faces) leads to a failure in learning generalized features and is likely to result in underfitting or catastrophic overfitting (in which the model only learns noise).
- **Transfer Learning Models (InceptionV1, ResNet50, MobileNetV2):** All models utilizing ImageNet pre-trained weights showed significantly superior performance, validating the power of transfer learning.
- **InceptionV1:** Achieved the **highest validation accuracy (82.72%)**. This suggests that the dense, 3×3 convolutional structure of VGG, coupled with the richly learned features from ImageNet, proved highly effective for capturing the subtle, high-frequency details necessary for face classification.

- **ResNet50:** Showed good accuracy (68.28%). Its architecture is robust and deep, benefiting from residual connections, but in this specific configuration, it was slightly outperformed by the VGG-19 transfer learning approach.
- **MobileNetV2:** Provided moderate accuracy (59.44%), which is a strong result considering its highly efficient, lightweight design focused on minimal computation.

Highlighting the Pros and Cons of Each Architecture

Architecture	Pros	Cons
VGG-19 (Scratch)	Conceptually simple, high feature-capturing capacity.	**Extremely low accuracy (4.51%):** Requires massive datasets; computationally expensive.
InceptionV1	**Highest Accuracy (82.72%)**; leverages rich features from ImageNet; strong capacity for feature extraction.	High memory footprint; slower inference compared to MobileNetV2.
ResNet50 (Transfer)	Excellent performance; residual connections stabilize training for deep networks; better generalization potential.	Very deep, still computationally demanding compared to mobile networks.
MobileNetV2 (Transfer)	**Most efficient** (low latency, low parameter count); ideal for deployment on edge devices.	Lower peak accuracy than deeper, more complex models.

Why VGG-19 Pre-trained is the Best Model

The VGG-19 pre-trained model demonstrated the **best performance** (82.72% validation accuracy) on this face classification dataset.

Reasoning:

1. **Transfer Learning Synergy:** The foundational features (edges, corners, general object parts) learned by VGG-19 on the massive ImageNet dataset are highly transferable to face images.
2. **Feature Richness:** The sequential, uniform 3×3 convolutional blocks in VGG-19 result in an exceptionally rich feature representation space, allowing the

network to distinguish between the subtle facial characteristics that define the 31 individual classes.
3. **Training Stability:** By using pre-trained weights and fine-tuning, the VGG-19 model was able to achieve a high-performance state quickly and stably, avoiding the disastrous results of training from scratch.

7. Conclusion

This comparative analysis successfully demonstrated the power of deep learning for multi-class face classification, confirming that the strategic application of **transfer learning significantly improves performance** over training complex models from scratch, especially on moderately sized datasets.

The **InceptionV1 architecture utilizing ImageNet pre-trained weights** emerged as the optimal model for this project, achieving the highest classification accuracy (82.72%). While ResNet50 offers excellent design principles and MobileNetV2 provides superior computational efficiency, VGG-19's capacity for fine-grained feature learning proved decisive for achieving peak accuracy in this identity classification task.

Practical Recommendation: For new face classification tasks where accuracy is the paramount concern and computational resources are sufficient, a pre-trained VGG-19 or ResNet model is recommended. If the solution must be deployed on a mobile or embedded system, MobileNetV2 offers the most viable and efficient high-accuracy alternative. Future work could involve integrating more specialized face recognition loss functions (e.g., ArcFace) to further enhance feature discriminability.