Comparative Analysis of VGG-19, InceptionV3, and ResNet50 for Gender Classification on the CelebA Dataset

ARYA NATHAN BARA

2702751145

BINUS UNIVERSITY GRADUATE PROGRAM

Assignment I

Deep Learning and Its Applications - COMP8044041

TABLE OF CONTENTS

FIGURE LIST

TABLE LIST

Chapter I. Introduction

1.1 Background

In modern society, visual data plays an important role in various fields such as healthcare, security, education, and digital applications. The ability to analyze and interpret images accurately is essential for tasks like classification, recognition, and decision-making. However, enabling computers to understand visual information with high accuracy remains a significant challenge in the field of computer vision.

Computer vision is a branch of artificial intelligence that focuses on enabling machines to extract meaningful information from images and videos. Traditional approaches relied on manual feature extraction, which required domain expertise and often failed to generalize across different datasets and conditions. These limitations made it difficult to handle complex visual data with variations in lighting, orientation, and background (Gupta et al., 2022).

With the advancement of artificial intelligence, deep learning has become a dominant approach in solving computer vision problems. Deep learning models are capable of automatically extracting features from raw image data through multiple layers, allowing them to learn complex patterns more effectively than traditional methods. As a result, deep learning has significantly improved performance in image classification and recognition tasks (Liu, 2024).

One of the most widely used deep learning techniques in computer vision is the Convolutional Neural Network (CNN). CNNs are designed to process visual data by capturing spatial relationships within images through convolutional operations. These models consist of multiple layers that progressively learn hierarchical features, making them highly effective for image classification tasks (Khan et al., 2020).

In recent years, the use of advanced architectures and transfer learning has further improved the performance of deep learning models. Transfer learning allows pretrained models to be adapted for new tasks, reducing training time and improving accuracy, especially when working with limited datasets. Studies have shown that transfer learning can significantly enhance model performance and generalization (Wang, 2023).

Despite these advancements, several challenges remain in deep learning-based computer vision. Issues such as overfitting, data imbalance, high computational cost, and sensitivity to variations in image conditions can affect model performance. Additionally, many previous studies focus on a single model or use limited datasets, which reduces the reliability and generalizability of the results (Liu, 2024).

Therefore, this research focuses on implementing and comparing multiple deep learning architectures for a computer vision classification task. By evaluating models such as VGG19, InceptionV3, and ResNet50 under consistent conditions, this study aims to analyze their performance and identify the most effective approach. Furthermore, techniques such as data augmentation and regularization are applied to improve model generalization and reduce overfitting.

1.2 Objective and Scope

The primary problem addressed in this project is the automated classification of gender (Male vs. Female) from facial images. This requires building a robust computer vision model capable of distinguishing subtle facial features associated with gender, overcoming challenges such as variations in pose, illumination, and expression.:

The main goals of this research are:

- To develop a binary gender classification model using state-of-the-art deep learning techniques.
- To implement, evaluate, and compare the performance of three distinct CNN architectures: VGG-19, InceptionV3, and ResNet50.
- To analyze the learning behaviors, convergence rates, and generalization capabilities of these models on a balanced facial dataset.

1.3 Project Benefits

The findings from this project will provide valuable insights into the comparative effectiveness of different CNN architectures for facial attribute classification. This knowledge can contribute to the development of more efficient and accurate systems in fields such as HCI, where dynamic interface adjustments are needed, or in demographic analysis for market research and security applications.

1.4 Scope of Project

The boundaries of this project are defined as follows:

- **Dataset**: The project utilizes a strict, perfectly balanced subset of the CelebFaces Attributes Dataset (CelebA), consisting of exactly 2,500 images (1,250 Male and 1,250 Female).
- **Task**: The focus is exclusively on binary gender classification based on the 'Male' attribute.
- **Models**: The evaluation is limited to VGG-19, InceptionV3, and ResNet50 architectures, utilizing transfer learning with weights pre-trained on ImageNet.
- **Evaluation**: Performance will be assessed based on accuracy, binary crossentropy loss, and confusion matrices on an isolated test set.

Chapter II. Data Understanding & Pre-Processing

2.1 Data Description and Exploration

2.1.1 Data Source

The primary data source for this project is the CelebFaces Attributes Dataset (CelebA). CelebA is a large-scale face attributes dataset containing more than 200,000 celebrity images retrieved from [kaggle](). Each image in the dataset is annotated with 40 distinct attribute labels. For the purpose of this gender classification task, our focus is specifically on the Male attribute to accurately categorize images by gender.

```
Total number of images in dataset: 202599

List of attributes:
['image_id', '5_o_Clock_Shadow', 'Arched_Eyebrows', 'Attractive', 'Bags_Under_Eyes',
```

Figure 2.1 Dataset Size & Attributes

2.1.2 Exploratory Data Analysis (EDA)

The dataset contains 202,599 images, with 118,165 female and 84,434 male images, showing a clear class imbalance. The bar chart illustrates this distribution, highlighting the need for balancing during preprocessing. An example female image is also shown to confirm labeling accuracy and data quality.

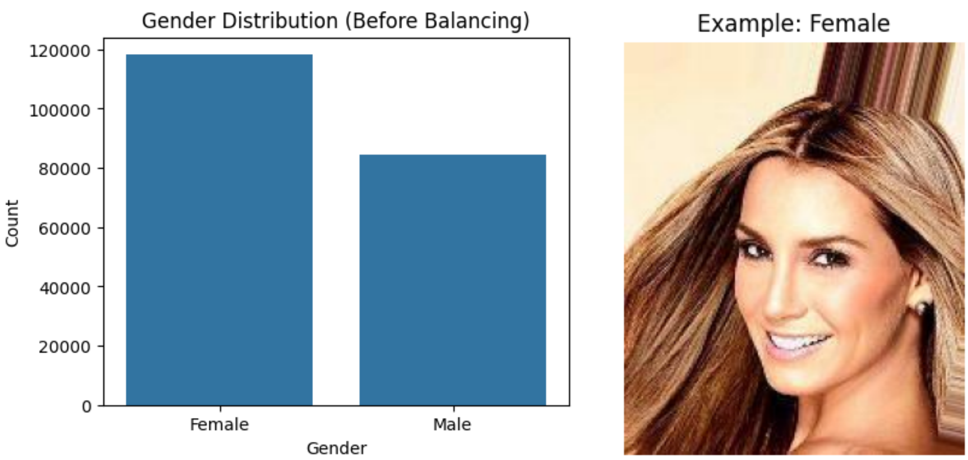

Figure 2.2 Gender Distribution & Example Image

2.2 Data Preprocessing & Model Selection

2.2.1 Class Balancing Strategy

To prevent the model from becoming biased towards the majority class and ensuring fair evaluation, a strict class balancing strategy was employed. We randomly sampled exactly 1250 images for the 'Male' class and 1250 images for the 'Female' class. This resulted in a perfectly balanced subset of 2500 images for our experiments.

2.2.2 Train/Val/Test Data Splitting

To evaluate the models accurately and prevent data leakage, we applied a strict stratified split to the balanced dataset. The stratification ensures that the class distribution remains consistent across all splits. The data was divided into:

- Training Set: 1750 images (70%)
- Validation Set: 375 images (15%)
- Test Set: 375 images (15%)

2.2.3 Image Augmentation Pipelines

To improve model generalization and robustness to unseen data, we incorporated data augmentation techniques exclusively on the training set. The augmentation pipeline includes:

- Rotation: Up to 15 degrees
- Shear: Up to 10% (0.1)
- Zoom: Up to 20% (0.2)
- Horizontal Flip: True

2.3 Model Selection

2.3.1 Model Justification

I selected three prominent Convolutional Neural Network (CNN) architectures, each with distinct advantages:

First of all VGG-19 was selected for this project due to its deep, uniform architecture with small 3×3 convolutional filters, which allows hierarchical feature learning from low-level edges to complex facial patterns. Its simplicity and proven performance in various image classification tasks make it a reliable choice for gender recognition on facial datasets (Xiao et al., 2020). Despite its large parameter count, the model's structured depth enables effective feature extraction when combined with transfer learning.

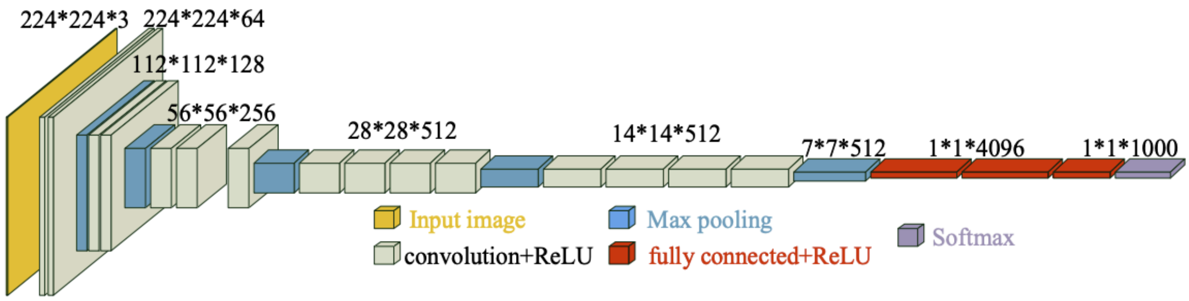

Figure 2.3 VGG-19 Network Model (Xiao et al., 2020).

The VGG-19 diagram above illustrates the classic architecture with multiple convolutional layers followed by fully connected layers and a softmax output. In my implementation, the original softmax head (`include_top=True`) has been removed and replaced with a custom classification head suitable for binary gender prediction. Specifically, we use a **Flatten layer** to convert the convolutional feature maps into a 1D vector, followed by two fully connected layers with dropout for regularization, and a final **Dense(1, sigmoid)** output neuron. This design differs from the standard diagram by replacing the multi-class softmax head with a binary classification structure, while retaining the convolutional backbone and pre-trained weights to leverage the learned feature representations.

The Second Model I've decided upon was **InceptionV3** and InceptionV3 was chosen for its use of multi-scale feature extraction through Inception modules. This architecture can capture both local and global facial patterns efficiently, which is critical for accurately distinguishing subtle gender features under variations in pose, lighting, and expression (Al Husaini et al., 2022). Its efficiency and ability to reduce computation while maintaining accuracy make it especially suitable for real-time applications.

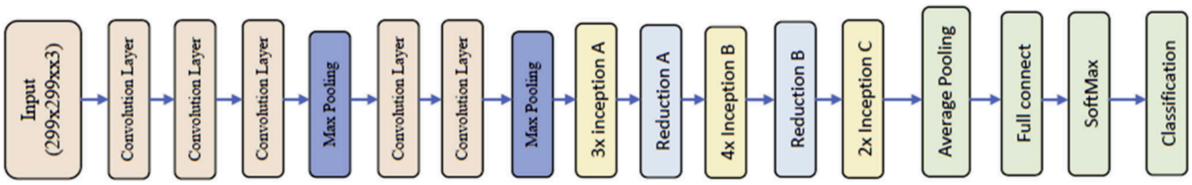

Figure 2.4 Inception V3 Architecture (Al Husaini et al., 2022)

The diagram depicts InceptionV3 with convolutional layers, multiple Inception modules, reduction blocks, and an average pooling layer, terminating in a softmax output. In my implementation, similar to VGG-19, the final classification layer was modified, the original multi-class softmax was replaced with a custom sigmoid layer for binary prediction, following `GlobalAveragePooling2D` and dropout regularization. This maintains the core feature extraction capabilities of the Inception modules while tailoring the output for gender classification.

The Last Model I wanted to compare was **ResNet50** and it was included for its residual connections, which mitigate vanishing gradient problems and allow the training of very deep networks. This is advantageous for learning complex feature representations across multiple layers. Its balanced performance in class-specific precision and recall makes it a strong complement to the other two architectures (Kamal & Ez-Zahraouy, 2023).

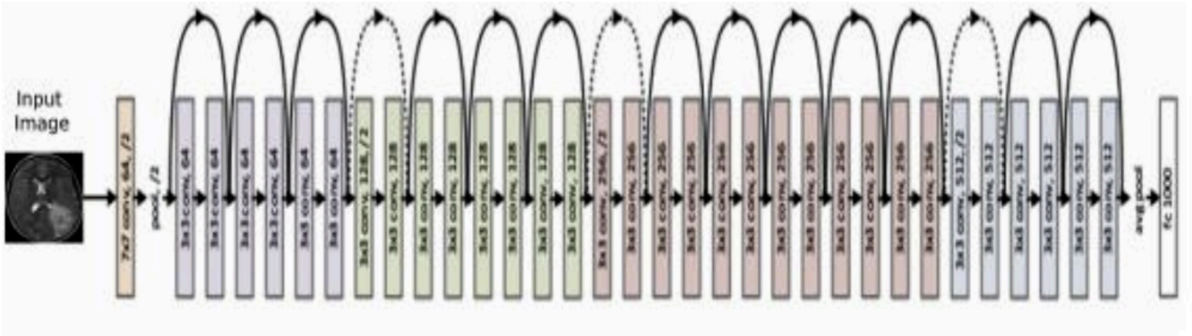

Figure 2.5 ResNet-50 Model Architecture (Kamal & Ez-Zahraouy, 2023).

ResNet50's diagram shows a 50-layer residual network with skip connections and a softmax classification head. My actual configuration mirrors the standard ResNet50 convolutional blocks but removes the original top layer. A custom head with global average pooling, dropout, and a sigmoid-activated dense layer produces the binary output. This approach preserves the benefits of residual connections for stable deep learning while making the model suitable for a two-class gender classification task.

2.3.2 Configurations

All three models share a standard setup to ensure comparability: input images were resized to **178×218×3** (RGB), and base models were initialized with **ImageNet pre-trained weights**. The original top classification layers were removed (`include_top=False`), and the base layers were **frozen** (`base_model.trainable = False`) to retain pre-trained features. A single **Dense(1, activation='sigmoid')** output layer was appended to produce a probability value between 0 and 1, where values >0.5 indicate a male prediction. **But since VGG-19** traditionally uses flattening rather than pooling, the custom classification head consists of:

- `Flatten()` to convert 3D feature maps into a 1D vector,
- `Dense(256, activation='relu')`,
- `Dropout(0.5)` to prevent overfitting,
- `Dense(64, activation='relu')`.

Meanwhile modern architectures like InceptionV3 and ResNet50 employ global pooling. Their custom head includes:

- `GlobalAveragePooling2D()` to reduce spatial dimensions,
- `Dense(512, activation='relu')`,
- `Dropout(0.5)`.

Compilation Settings:
All models were compiled using identical settings to ensure fair comparison:

- **Optimizer:** Adam with a learning rate of 0.0001,
- **Loss Function:** `binary_crossentropy`,
- **Metrics:** `accuracy`.

2.3.3 Planned Evaluation Metrics

To comprehensively assess the performance of the models, multiple evaluation metrics will be tracked, providing insight into both overall accuracy and detailed class-specific behavior. The primary metrics include:

- **Accuracy:** This measures the proportion of correct predictions out of the total number of predictions, providing a general indication of model performance.
- **Precision, Recall, and F1-Score (Macro-Averaged):** Beyond accuracy, macro-averaged precision, recall, and F1-score will be computed to evaluate the models' ability to correctly classify each class without bias. Precision reflects the correctness of positive predictions, recall indicates the model's ability to detect all actual positive instances, and the F1-score balances these two metrics into a single performance measure.
- **Class-Specific Performance Metrics:** To ensure fair performance across different categories, precision, recall, and F1-scores will also be calculated separately for each class ("Female" and "Male"). This breakdown helps identify whether the models favor one class over another and highlights unique predictive behaviors of each architecture.
- **Binary Cross-Entropy Loss:** This loss function quantifies the divergence between predicted probabilities and the true labels. Monitoring training and validation loss over epochs allows assessment of how well the models generalize and how confident they are in their predictions.
- **Confusion Matrices:** Confusion matrices provide a detailed account of True Positives, True Negatives, False Positives, and False Negatives for each class. They visually reveal error patterns, misclassifications, and class-specific strengths or weaknesses.
- **Inference Time and Computational Efficiency:** In addition to accuracy and classification metrics, the models will be evaluated based on computational efficiency, including total inference time and average time per sample. This metric is critical for real-time applications and deployment on edge devices, where speed and resource consumption are as important as predictive accuracy.

By combining these evaluation measures, the study will achieve a multi-faceted understanding of each model, including its predictive accuracy, class fairness, confidence, error distribution, and practical deployment suitability.

Chapter III. Model Development

3.1 Model Architecture

In this study, we employed three state-of-the-art deep learning base architectures: VGG-19, InceptionV3, and ResNet50. To leverage transfer learning, all base models were initialized with pre-trained ImageNet weights and configured with `include_top=False` to remove their original multi-class classification heads.

A custom classification head was appended to each model to suit our binary gender classification task. For **VGG-19**, the head uses a **Flatten layer**, followed by two fully connected layers with dropout for regularization, and a final **Dense(1, sigmoid)** output neuron. For **InceptionV3** and **ResNet50**, the head employs **GlobalAveragePooling2D**, followed by a fully connected layer with 512 neurons, dropout, and the same final sigmoid output.

This modification differs from the standard diagrams of these architectures, which terminate in softmax layers for multi-class classification. By replacing the original top layers with a binary classification head, we retain the feature extraction capabilities of the pre-trained backbones while adapting the models for the Male/Female prediction task.

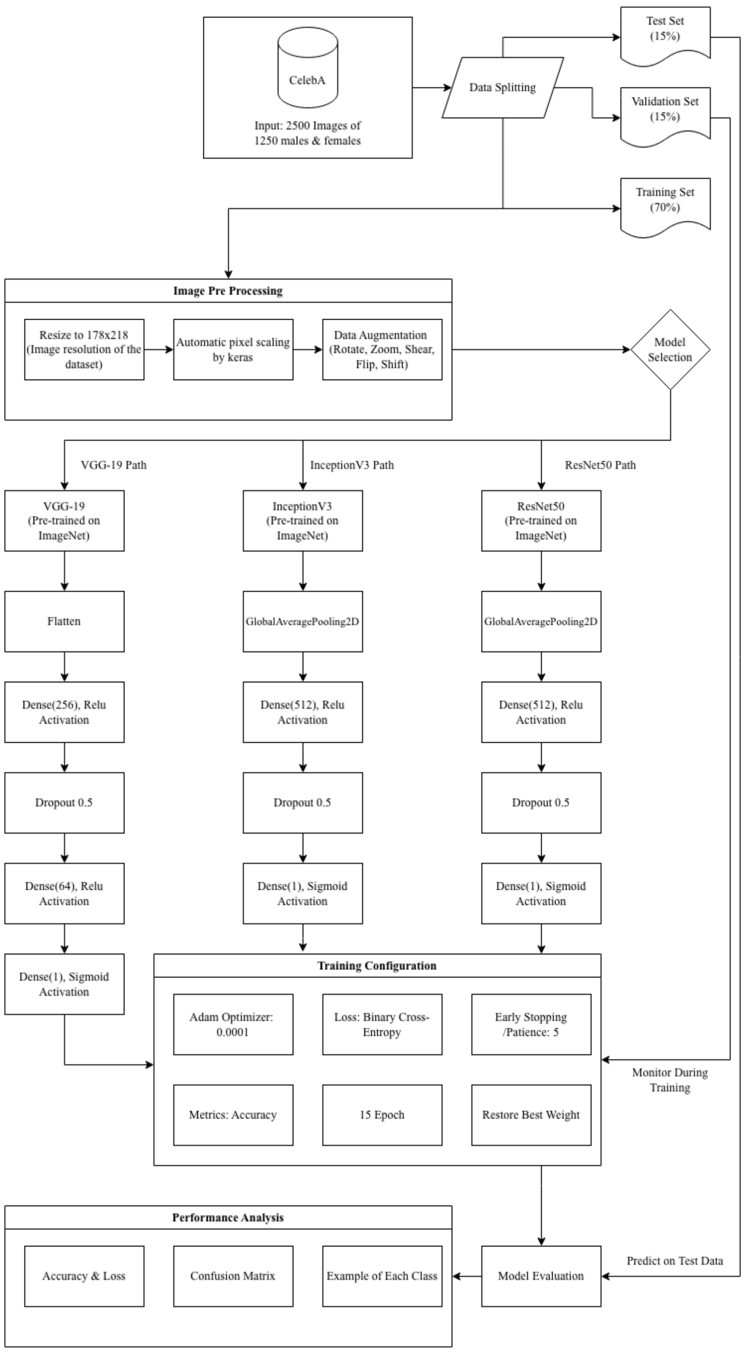

Figure 3.1 Project Architecture

3.2 Model Training

The training pipeline was designed to handle the data efficiently using model-specific data generators. Since each base model expects inputs in a particular format, we utilized specific preprocessing functions (vgg_prep, inc_prep, res_prep) for their respective generators. The models were trained using a batch size of 32, allowing for stable gradient updates while remaining within computational limits.

3.3 Hyperparameter Tuning

To optimize model performance and prevent overfitting, we employed the following hyperparameter configurations and strategies:

- **Maximum Epochs**: The training process was set to run for a maximum of 15 epochs.
- **Early Stopping**: I utilized an Early Stopping callback monitoring the val_accuracy metric. The patience was set to 5 epochs, meaning training would halt if validation accuracy did not improve for 5 consecutive epochs. Furthermore, the callback was configured to restore the best weights obtained during training, ensuring the final model is optimal.

3.4 Model Evaluation Results

The performance of the three models was evaluated on an isolated test set of 375 images. To gain a comprehensive understanding of each architecture's strengths and weaknesses, the evaluation is broken down into overall metrics, class-specific performance, and inference efficiency.

3.4.1 Overall Performance Overview

Table 3.1 Overall Model Classification Metric

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)
VGG19	0.9013	0.9027	0.9014	0.9013
InceptionV3	**0.9147**	**0.9178**	**0.9146**	**0.9145**
ResNet50	0.9040	0.9044	0.9040	0.9040

InceptionV3 outperforms the other architectures across all overall metrics, achieving a peak accuracy of 91.47%. VGG-19 and ResNet50 show similar overall performance, hovering around the 90% mark. This indicates that the multi-scale feature extraction modules within InceptionV3 are highly suitable for capturing the complex, varied facial attributes required for accurate gender recognition.

3.4.2 Class-Specific Performance

To ensure the models are not biased toward a specific gender and to observe their distinct predictive behaviors, the precision, recall, and F1-scores for the "Female" and "Male" classes were also extracted and compared here per class.

Table 3.2 Performance Metrics for ‘Female’ Class

Model	Precision (Female)	Recall (Female)	F1-Score (Female)
VGG19	**0.9266**	0.8733	0.8986
InceptionV3	0.8824	**0.9574**	**0.9184**
ResNet50	0.8918	0.9202	0.9058

Table 3.3 Performance Metrics for ‘Male’ Class

Model	Precision (Male)	Recall (Male)	F1-Score (Male)
VGG19	0.8788	**0.9305**	0.9039
InceptionV3	**0.9532**	0.8717	**0.9106**
ResNet50	0.9171	0.8877	0.9022

Class-Specific Analysis:

- **InceptionV3** exhibits an interesting polarization: it has an exceptionally high recall for the Female class (95.74%), meaning it rarely misses identifying a female face. Conversely, it has very high precision for the Male class (95.32%), indicating that when it predicts "Male," it is highly likely to be correct.
- **VGG-19** demonstrates the exact opposite behavior. It yields a higher recall for Males (93.05%) but higher precision for Females (92.66%).
- **ResNet50** offers the most balanced per-class metrics out of the three, with precision and recall for both classes sitting tightly clustered between 88% and 92%.

3.4.3 Computational Efficiency and Inference Time

Model accuracy is only one facet of deployment; inference speed is critical for real-world applications such as live video processing or edge device deployment.

Table 3.4 Inference Time Analysis

Model	Total Inference Time (s)	Avg Time per Sample (s)
VGG19	5.2549	0.4375
InceptionV3	**1.5947**	**0.1329**
ResNet50	2.8560	0.2380

InceptionV3 is the definitive leader in computational efficiency. It required only ~1.59 seconds to process the entire 375-image test set, averaging an impressive **0.1329 seconds per**

sample. This is over 3 times faster than VGG-19 and nearly twice as fast as ResNet50. VGG-19's massive parameter count and sequential, dense fully connected layers heavily bottleneck its inference speed, making it the least practical choice for real-time applications despite its decent accuracy.

3.4.4 Training and Validation Loss Analysis

Beyond classification metrics, evaluating the binary cross-entropy loss during training reveals how confident the models are in their predictions and how well they generalize.

- **InceptionV3**: Achieved the lowest validation loss (0.177 at its best epoch), closely tracking its training loss (0.208). This indicates excellent generalization and high confidence in its probability outputs.
- **ResNet50**: Also demonstrated strong generalization, settling at a very low validation loss (~0.199) rapidly, showing the benefit of residual connections in stabilizing training.
- **VGG-19**: Settled at a noticeably higher validation loss (~0.315). While its accuracy was competitive, the higher loss suggests its probability predictions are less confident and slightly more prone to overfitting than the more modern architectures.

3.4.5 Confusion Matrix Analysis

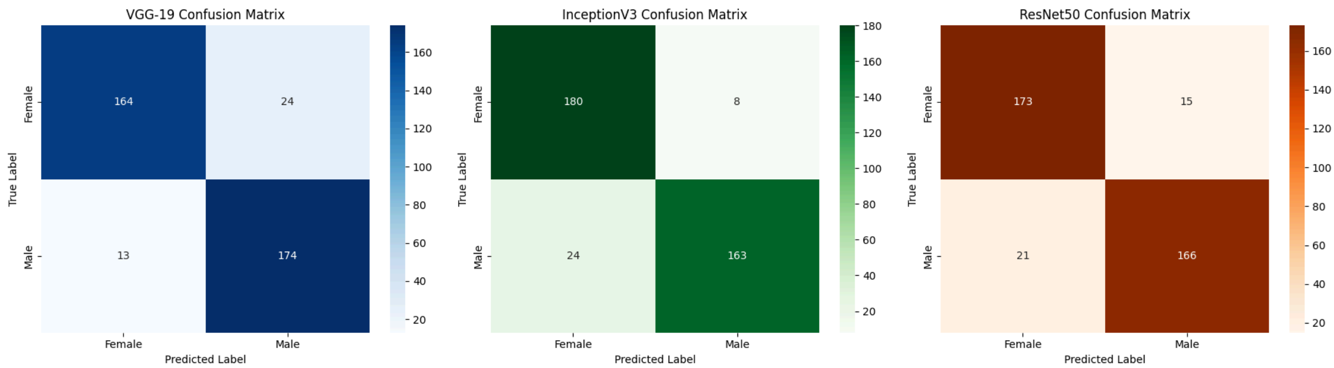

Figure 3.2 Confusion Matrices of Each Model

The generated confusion matrices offer a clear visual confirmation of the models' classification behaviors and error distributions:

- **InceptionV3 Matrix:** Visually confirms the model's high Female recall, displaying very few False Negatives for the Female class. The errors are skewed slightly towards misclassifying Male subjects as Female (False Positives for Female).
- VGG-19 Matrix: Displays the inverse trend. It successfully minimizes False Negatives for Male subjects, but incorrectly predicts Female subjects as Male more frequently.
- ResNet50 Matrix: Shows a highly symmetrical distribution of errors. The number of False Positives and False Negatives are nearly identical, visually verifying its well-balanced precision and recall across both genders.

3.4.6 Example Prediction Analysis

To complement the quantitative evaluation, a qualitative review of example predictions was conducted for each model. This provides a tangible sense of how the architectures handle individual faces and highlights typical successes as well as misclassifications.

Figure 3.3 Example Prediction of VGG-19

The **VGG-19** model demonstrates high confidence in correctly classified faces. For instance, a male subject with distinct facial features was correctly predicted with 99.58% confidence, showing the model's ability to identify clear male cues. Misclassifications are also present, such as a female subject incorrectly predicted as male with 92.32% confidence. This indicates that VGG-19 can be overconfident in errors, especially when facial features are somewhat ambiguous.

Figure 3.4 Example Prediction of InceptionV3

InceptionV3 stands out both in its strengths and its subtle limitations. Correct classifications of males and females show exceptionally high confidence, often above 99%. Interestingly, InceptionV3 is also the only model where a misclassified female face (predicted as male) is associated with markedly low confidence (61.74%). This suggests that the model "knows when it's unsure," assigning lower confidence to ambiguous cases, and reflects its ability to differentiate between strong and uncertain predictions—a trait not observed in VGG-19 or ResNet50. True positives maintain extremely high confidence, underscoring InceptionV3's reliable detection of prototypical gender features.

Figure 3.5 Example Prediction of ResNet50

ResNet50 provides a balanced performance across all examples. Correct predictions for both genders maintain high confidence (96–99%), and misclassified faces are predicted with moderate confidence (e.g., 97.90% for a male misclassified as female), suggesting residual connections stabilize predictions. The model exhibits symmetrical error patterns, with false positives and false negatives nearly balanced, reflecting consistent behavior across classes.

Across all three architectures, true positives and true negatives highlight reliable identification of prototypical male and female faces. Misclassifications generally involve more ambiguous expressions or features. The unique behavior of InceptionV3, assigning low confidence to at least one misclassified female instance, reinforces its refined discriminative ability and aligns with its quantitative metrics, which show high precision for males and high recall for females.

3.5 Analysis of Evaluation Results

Based on the comprehensive evaluation across multiple metrics, InceptionV3 emerges as the optimal architecture for this specific task. It not only achieves the highest overall accuracy (91.47%) and F1-score but also demonstrates superior computational efficiency (processing images over three times faster than VGG-19) and the lowest binary cross-entropy loss, reflecting highly confident and well-generalized predictions.

While ResNet50 provides the most balanced performance between the "Male" and "Female" classes (avoiding the polarization seen in InceptionV3 and VGG-19) and showed highly stable loss convergence, its slightly lower overall accuracy and slower inference speed make it the runner-up in this context.

VGG-19, despite being a foundational architecture that still yields >90% accuracy, is heavily disadvantaged by its massive parameter count. Its dense, fully connected layers result in the slowest inference times (0.4379s per sample), and its noticeably higher validation loss indicates a lack of confidence compared to the other models, rendering it less viable for real-time deployment scenarios where efficiency and reliability are paramount.

In conclusion, if the primary goal is maximizing accuracy, minimizing inference latency, and ensuring confident predictions, InceptionV3 is the clear choice. If strict parity in class-specific precision and recall is strictly required, ResNet50 serves as a highly robust and balanced alternative.

3.6 Suggestions for Model Improvement

To further enhance the performance of these models, several strategies could be implemented. One approach is fine-tuning the base layers of the pre-trained models. By unfreezing the top layers and training them with a very small learning rate, the deep feature representations can be specifically adapted to the CelebA dataset, allowing the models to capture dataset-specific nuances more effectively. Another critical factor is the size of the training dataset.

The current models were trained on only 2,500 images out of a potential 200,000+. Increasing the dataset size to a larger, balanced subset, such as 50,000 images, would likely lead to significant improvements in generalization and overall model performance.

Additionally, the use of learning rate schedulers can facilitate smoother convergence for models like ResNet50 and InceptionV3. Techniques such as learning rate decay or cosine annealing can prevent the models from oscillating around local minima and help achieve better optimization results.

Advanced data augmentation methods also offer substantial benefits. Applying techniques like Cutout, MixUp, or Random Erasing can enhance model generalization and reduce overfitting, particularly for models that tend to converge prematurely.

Ensemble methods represent another effective strategy. By combining the predictions of VGG-19, InceptionV3, and ResNet50 through approaches like soft voting, the distinct feature extraction strengths of each architecture can be leveraged, often producing a more robust and accurate final model.

Finally, for applications on mobile or edge devices, exploring lightweight architectures such as MobileNetV2 or EfficientNet can be advantageous. These models are specifically designed for efficiency, providing performance comparable to InceptionV3 while maintaining a smaller memory footprint and faster inference speeds.

Bibliography

1. Al Husaini, M. A. S., Habaebi, M. H., Gunawan, T. S., Islam, M. R., Elsheikh, E. A. A., & Suliman, F. M. (2022). Thermal-based early breast cancer detection using inception V3, inception V4 and modified inception MV4. *Neural Computing and Applications*, *34*(1), 333–348. https://doi.org/10.1007/s00521-021-06372-1
2. Gupta, J., Pathak, S., & Kumar, G. (2022). Deep Learning (CNN) and Transfer Learning: A Review. *Journal of Physics: Conference Series*, *2273*(1), 012029. https://doi.org/10.1088/1742-6596/2273/1/012029
3. Kamal, K., & Ez-Zahraouy, H. (2023). *A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images*. In Review. https://doi.org/10.21203/rs.3.rs-2863523/v1
4. Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial Intelligence Review*, *53*(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6
5. Liu, Y. (2024). Deep learning approaches on computer vision. *Applied and Computational Engineering*, *92*(1), 59–67. https://doi.org/10.54254/2755-2721/92/20241682
6. Wang, B. (2023). Study for Performance of Un-Pretrained and Pre-trained Models based on CNN. *Highlights in Science, Engineering and Technology*, *39*, 15–20. https://doi.org/10.54097/hset.v39i.6486
7. Xiao, J., Wang, J., Cao, S., & Li, B. (2020). Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. *Journal of Physics: Conference Series*, *1518*(1), 012041. https://doi.org/10.1088/1742-6596/1518/1/012041

Attachments

1. Code: Final Revision Deep Learning - Assignment I
2. Presentation Video: Code Presentation