CS412: Machine Learning

Homework 3: Transfer Learning with VGG-16

Name: Kerem Tufan
ID: 00032554
Date: December 7, 2025

Colab Notebook Link:
Click here to access the Notebook

Sabancı University

1 Introduction

In this homework, the goal is to perform binary classification on facial images to detect whether a person is smiling or not. A subset of the CelebA dataset is utilized, and Transfer Learning techniques are employed using a pre-trained VGG-16 model.

The problem is approached as a binary classification task where the pre-trained feature extractor is adapted to the specific dataset. Two different strategies are evaluated: freezing the convolutional base versus fine-tuning the last convolutional block. This report details the methodology, experimental results, and a discussion on the impact of data augmentation and fine-tuning strategies.

2 Method

2.1 Dataset and Preprocessing

The dataset consists of 30,000 celebrity face images with the "Smiling" attribute. Analysis of the dataset distribution showed a balanced class ratio:

- **Smiling (1):** 14,259 images (47.53%)
- **Not Smiling (0):** 15,741 images (52.47%)

Experimental setup details:

- **Data Split:** The dataset is split into 80% training, 10% validation, and 10% test sets.
- **Data Augmentation:** For the training set, random horizontal flips ($p = 0.5$) and light color jitter ($brightness = 0.2$, $contrast = 0.2$, $saturation = 0.2$, $hue = 0.1$) were applied to increase model robustness against variations.
- **Preprocessing:** All images were resized to 224×224 and normalized using mean=[0.5, 0.5, 0.5] and std=[0.5, 0.5, 0.5] to scale pixel values between -1 and 1.

2.2 Model Architecture and Strategies

The VGG-16 architecture pretrained on ImageNet was utilized. The original classification head was replaced with a binary output layer (Linear layer with 1 output) suitable for the task. Two strategies were evaluated:

1. **Strategy 1 (Freeze All):** All convolutional layers were frozen; only the classifier head was trained.
2. **Strategy 2 (Fine-Tune):** The weights of the entire network were frozen except for the last convolutional block (Block 5) and the classifier head.

Training was performed for 10 epochs using the Binary Cross Entropy with Logits Loss and the Adam optimizer.

3 Results

Training Dynamics: As shown in Figure 1, Strategy 2 achieved its peak validation accuracy ($\approx 92.5\%$) around Epoch 3-4. After this point, the Validation Loss began to increase (forming a U-shape), indicating the onset of overfitting. In contrast, Strategy 1 remained stable but stuck at a lower performance plateau ($\approx 86\%$), proving that the frozen feature extractor was insufficient for capturing the fine-grained details of smiling.

Figure 1: Comparison of Validation Loss and Accuracy for Strategy 1 (Freeze) and Strategy 2 (Fine-Tune).

3.1 Model Comparison Table

The following table summarizes the performance metrics and training duration for both strategies.

Model Setting	Final Val Acc	Test Acc	Test Loss	Duration (min)
Strategy 1 (Freeze All)	0.8580	0.8540	0.4163	57.72
Strategy 2 (Fine-Tune)	0.9173	0.9063	0.3122	59.98

Table 1: Performance comparison of different training strategies.

Based on the validation set performance, **Strategy 2 (Fine-Tune)** was selected as the best model, achieving significantly higher accuracy and lower loss compared to the frozen feature extractor approach.

3.2 Test Set Evaluation

The confusion matrix for the best performing model (Strategy 2) on the test set is shown in Figure 2.

The matrix demonstrates the model's robust performance with a **Test Accuracy of 90.63%**. Out of 3,000 test samples, the model correctly classified 1479 "Not Smiling" and

1240 "Smiling" faces. Notably, the model produced fewer False Positives (95) compared to False Negatives (186), indicating a slight tendency to predict the negative class ("Not Smiling") when the facial expression is ambiguous.

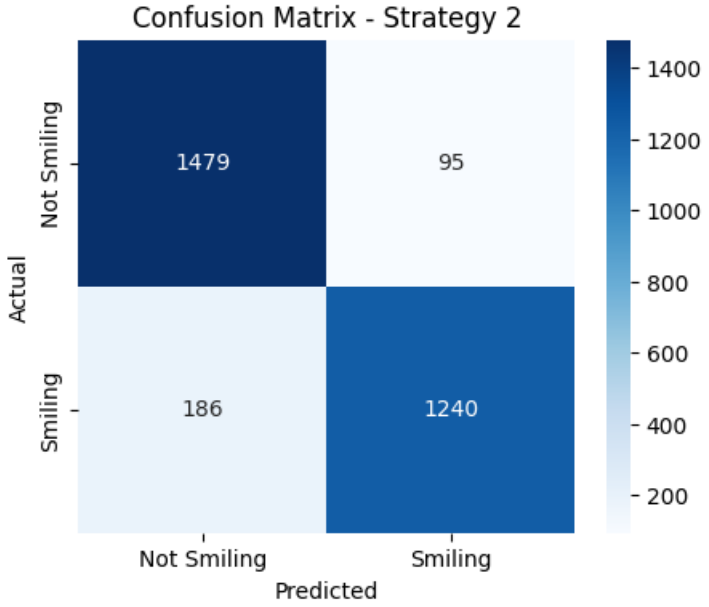
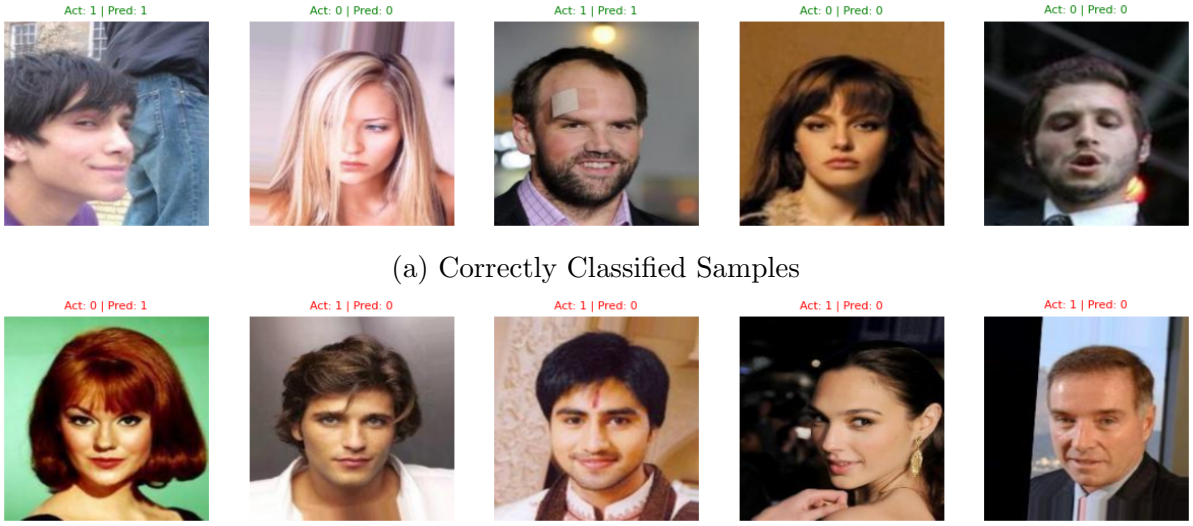


Figure 2: Confusion Matrix of the best model (Strategy 2) on the Test Set.

3.3 Qualitative Results

Below are examples of correct and incorrect classifications by the model.



(a) Correctly Classified Samples

(b) Incorrectly Classified Samples

Figure 3: Qualitative analysis of model predictions.

Observation on Correct Samples: The model demonstrates robustness against occlusions (e.g., the bandage in Image 3 of correct samples) and successfully distinguishes between "open mouth" and "smiling" (e.g., Image 5 of correct samples, where the subject is singing/talking but correctly classified as not smiling).

4 Discussion

4.1 Impact of Fine-Tuning vs. Transfer Learning

A comparison between Strategy 1 (Freeze All) and Strategy 2 (Fine-Tune) reveals the following:

- **Feature Adaptation:** Strategy 1 uses generic features learned from ImageNet. While effective, these features might not be perfectly aligned with the nuances of smiling faces. Strategy 2, by unfreezing the last block, allows the model to adapt high-level features (such as mouth shape and eye crinkling) specifically for this task.
- **Performance:** As observed in Table 1, fine-tuning generally yields higher accuracy (90.63% vs 85.40%) because it specializes the feature extractor. Strategy 2 also achieved a much lower test loss (0.3122) compared to Strategy 1 (0.4163), indicating higher confidence in correct predictions.

4.2 Impact of Data Augmentation

To analyze the effect of data augmentation, the standard training (Freeze All with flips and color jitter) was compared against a non-augmented training session (Freeze All without augmentation).

- **Without Augmentation:** The model showed signs of overfitting. While the training accuracy increased, the Validation Loss diverged significantly (reaching ≈ 0.78), indicating that the model was memorizing the training data and becoming "over-confident" in its wrong predictions on unseen data.
- **With Augmentation:** The Validation Loss remained much lower and stable (≈ 0.40). This demonstrates that augmentation acts as a regularizer, forcing the model to learn robust features (invariance to color/orientation) rather than memorizing pixel patterns, thus improving generalization.

4.3 Error Analysis

An analysis of the incorrectly classified samples (Figure 3b) highlights specific challenges for the model:

- **Ambiguity and Subjectivity:** Several misclassified images contain "smirks" or very subtle expressions (e.g., Image 1 in incorrect samples). The boundary between a neutral face and a slight smile is subjective. For instance, the model predicted a smile for the subject in Image 1 due to upturned mouth corners, despite the ground truth label being "Not Smiling".
- **Pose Variation:** Image 4 (incorrect samples) features a subject looking over her shoulder. This extreme head pose alters the geometric appearance of facial features. The model failed to detect the smile from this angle, predicting "Not Smiling".
- **Subtle Smiles:** Images 2 and 3 in the incorrect set feature subjects with closed-mouth smiles. The model struggled to distinguish these subtle expressions from neutral faces, resulting in False Negatives.