**Deep Learning for Perception**

**Project Report**



**Facial Recognition using Vision Transformer**

Submitted by:

21K-4836 Shayan Anwar

21K-4886 Ali Rafay

**1. Objective**

The objective of this project is to implement and compare facial recognition systems using **Vision Transformers (ViT)** and a **custom-built FaceNet model** on the Labeled Faces in the Wild (LFW) dataset. The specific goals are:

* Fine-tune a pre-trained ViT model and train a FaceNet architecture from scratch on the LFW dataset.
* Evaluate both models' performance on unseen data.
* Compare the effectiveness of modern transformer-based and traditional embedding-based approaches in facial recognition.

**2. Problem Statement**

Facial recognition is a key area in computer vision with applications in security, identity verification, and social media. The challenge is to develop models that perform well under real-world variations like lighting, pose, and expression.

The problem for this project is to:

* Train deep learning models that can accurately identify individuals from facial images in the LFW dataset.
* Ensure these models generalize well to varied conditions.

## 3. **Methodology**

### 3.1 Dataset

The LFW dataset contains over 13,000 labeled images of 5,749 individuals. Images are collected in unconstrained settings, introducing variation in lighting, pose, and background.

### 3.2 Model Selection

#### a. Vision Transformer (ViT)

* **Architecture**: ViT-B/16 from torchvision.models, pre-trained on ImageNet.
* **Modification**: The final classification head was replaced to output 5,749 classes, matching the number of individuals in the dataset.
* **Purpose**: Leverages attention mechanisms for global feature understanding.

**b. FaceNet (from scratch)**

* **Architecture**: A custom CNN-based FaceNet model inspired by the original Google paper.
  + The network uses convolutional blocks with batch normalization, followed by fully connected layers that output 128-dimensional face embeddings.
* **Loss Function**: Triplet Loss to enforce the similarity between same-person embeddings and separation between different individuals.
* **Embedding Training**: The model is trained to produce embeddings instead of class scores. Classification is then done using a nearest-neighbor classifier or cosine similarity on the learned embeddings.

### 3.3 Data Preprocessing

* **Resizing**: All images were resized to 224x224 pixels.
* **Normalization**: Images normalized using ImageNet mean/std for ViT, and custom normalization for FaceNet.
* **Augmentation**: Applied random horizontal flips, rotations, and random crops to improve generalization

### 3.4 Training Process

The model was trained using the following configuration:

1. **ViT**

* **Batch size**: 32
* **Optimizer**: AdamW with a learning rate of 1e-4 and weight decay of 0.01.
* **Learning Rate Scheduler**: A learning rate reduction strategy was employed using ReduceLROnPlateau based on the validation loss.
* **Loss Function**: Cross-entropy loss.
* **Training Time**: The model was trained for 5 epochs.

1. **FaceNet (Custom)**

* **Batch size**: 32
* **Optimizer**: AdamW with a learning rate of 1e-4 and weight decay of 0.01.
* **Loss Function**: Triplet Loss with semi-hard mining
* **Embedding Dimension**: 128
* **Training Strategy**
  + **Construct triplets: (anchor, positive, negative)**
  + **Use cosine similarity or L2 distance for evaluation**
* **Training Time**: The model was trained for 20 epochs ~5 hours of runtime.

The ViT model was trained on **Kaggle** using a GPU (NVIDIA T4 x2) for efficient computation however, with the data being so large it still took around 6 hours to run 5 epochs. The FaceNet model however ran faster. It was trained using CPU on google Colab.

### 3.5 Evaluation

#### ViT

* **Metrics**: Accuracy, confusion matrix
* **Evaluation**: Classification performance across 5,749 classes
* **Visualization**: Confusion matrix for misclassification insight

#### FaceNet

* **Metrics**:
  + **Verification accuracy**: Percentage of correct matches in face verification
  + **ROC-AUC**: Receiver Operating Characteristic to measure match quality
* **Evaluation**:
  + Embeddings were tested using cosine similarity and KNN-based classification.
  + Accuracy computed on a held-out test set using face pairs (genuine vs impostor).

## 4. **Results**

### **4.1 Confusion Matrix and Accuracy**

#### ViT

* A confusion matrix showed strong class-wise performance, though visually similar individuals caused some misclassifications.

#### FaceNet

* Achieved high verification accuracy on test pairs.
* ROC curve showed strong discriminative ability of embeddings with an AUC above 0.95.
* KNN-based classification also performed well, particularly on frequently occurring individuals.

(Further screenshots of results will be uploaded here)

**5. Conclusion**

This project implemented and compared two models for facial recognition using the LFW dataset: a fine-tuned ViT and a FaceNet-style model trained from scratch.

* **ViT** demonstrated strong classification accuracy and was easy to fine-tune using pre-trained weights.
* **FaceNet**, though more complex to train, provided robust and scalable face embeddings suitable for face verification tasks.

The results highlight the trade-offs between classification-based models (ViT) and embedding-based models (FaceNet), with both performing well in different scenarios.

**6. Future Work**

Future work could include:

* **Extended Datasets**: Try larger and more diverse datasets like VGGFace2 or MS-Celeb-1M.
* **Real-Time Applications**: Deploy FaceNet-based system in low-latency environments.
* **Model Improvements**: Test hybrid CNN-Transformer models or multi-task learning strategies.
* **Embedding Refinement**: Use hard negative mining or contrastive learning to further improve FaceNet embeddings.

**7. References**

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2020). *Attention is all you need*. In Advances in Neural Information Processing Systems (NeurIPS 2017).
2. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *CVPR*.
3. LFW Dataset for facial recognition.