**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 a facial recognition system using Vision Transformers (ViT) on the Labeled Faces in the Wild (LFW) dataset. The specific goals are:

* Fine-tune a pre-trained ViT model on the LFW dataset for facial recognition.
* Evaluate the model’s performance on unseen data.
* Achieve high accuracy in identifying individuals based on facial images.

**2. Problem Statement**

Facial recognition is an important area of computer vision with applications in security, personal identification, and social media. The challenge lies in creating models that can generalize well to real-world data, where lighting conditions, poses, and expressions can vary significantly.

The problem for this project is to:

* Train a deep learning model that can correctly identify individuals based on their facial images from the LFW dataset.
* Ensure that the model performs well even with variations in lighting, background, and pose.

## 3. ****Methodology****

### 3.1 **Dataset**

The LFW dataset contains 13,000 labeled images of 5,749 different individuals, making it a good benchmark for facial recognition tasks. The dataset is split into training and testing subsets, with images containing faces in different lighting conditions and poses.

### 3.2 **Model Selection**

We used **Vision Transformer (ViT)**, a cutting-edge model for image classification tasks. ViT, originally designed for natural language processing, treats images as a sequence of patches and processes them with a Transformer-based architecture.

* **Pre-trained ViT model**: We used the pre-trained ViT-B-16 model from torchvision, which was pre-trained on ImageNet.
* **Fine-tuning**: The last classification layer of the pre-trained model was replaced to match the number of classes in the LFW dataset (5,749).

### 3.3 **Data Preprocessing**

* **Resizing**: Images were resized to 224x224 pixels to match the input size required by the ViT model.
* **Normalization**: Images were normalized using the mean and standard deviation of the ImageNet dataset (since the model was pre-trained on ImageNet).
* **Augmentation**: Random horizontal flips, rotations, and crops were applied to increase the diversity of the training set and reduce overfitting.

### 3.4 **Training Process**

The model was trained using the following configuration:

* **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.

The 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.

### 3.5 **Evaluation**

After training, the model was evaluated on the validation set:

* **Metrics**: The model’s performance was evaluated using accuracy and loss. A confusion matrix was also computed to analyze the performance across different classes.
* **Validation Results**: The model's accuracy on the validation set was recorded, and a confusion matrix was generated for further insights into the performance.

## 4. ****Results****

### 4.1 **Confusion Matrix**

A confusion matrix was computed to examine how well the model identified individuals. The results indicate that while the model performs well, there are some misclassifications, likely due to similar appearances between some individuals or variations in the dataset.

(Further screenshots of results will be uploaded here)

**5. Conclusion**

In this project, we successfully fine-tuned a pre-trained Vision Transformer (ViT) model for facial recognition on the LFW dataset. The model achieved a high accuracy on the validation set, demonstrating that ViTs are effective for facial recognition tasks, even with complex variations in facial images.

The project illustrates the effectiveness of ViT in handling complex classification tasks like facial recognition, providing insights into the future potential of Transformer-based models in computer vision.

**6. Future Work**

Future work could include:

* **Extended Datasets**: Experimenting with other facial recognition datasets, such as VGGFace2 or CelebA, to further evaluate the model.
* **Model Enhancement**: Investigating larger variants of ViT (e.g., ViT-Large) or hybrid models combining CNNs and Transformers for better performance.
* **Real-Time Application**: Deploying the trained model in real-time applications for facial recognition in security systems or personal devices.

**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. LFW Dataset for facial recognition.