GENERATIVE AI & DL PROJECT REPORT

(PCCCS681)

SIMPLE FACE GENDER CLASSIFIER

SUBMITTED BY

JAYABROTA BANERJEE (2) – 12022002018078

ARITRA GHOSAL (23) – 12022002018036

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF

BACHELOR OF TECHNOLOGY IN

COMPUTER SCIENCE FOR BUSINESS SYSTEMS

SUPERVISED BY

DR SWARNENDU GHOSH

INSTITUTE OF ENGINEERING & MANAGEMENT

(UNDER MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY)

APRIL 2025

TABLE OF CONTENTS

INTRODUCTION

This project is about creating a model that guesses the gender of a person from a single image of the person's face. Initially a face is detected then it is classified into male or female.

DATASET

Dataset was generated for this project using a separate model made from scratch. It randomly places features around places where features tend to occur statistically for males and females. The model is trained using synthetic data. To use custom data, place the image files in the "data" directory, divided into "test" and "train" subdirectories, further divided into "male" and "female".

MODEL

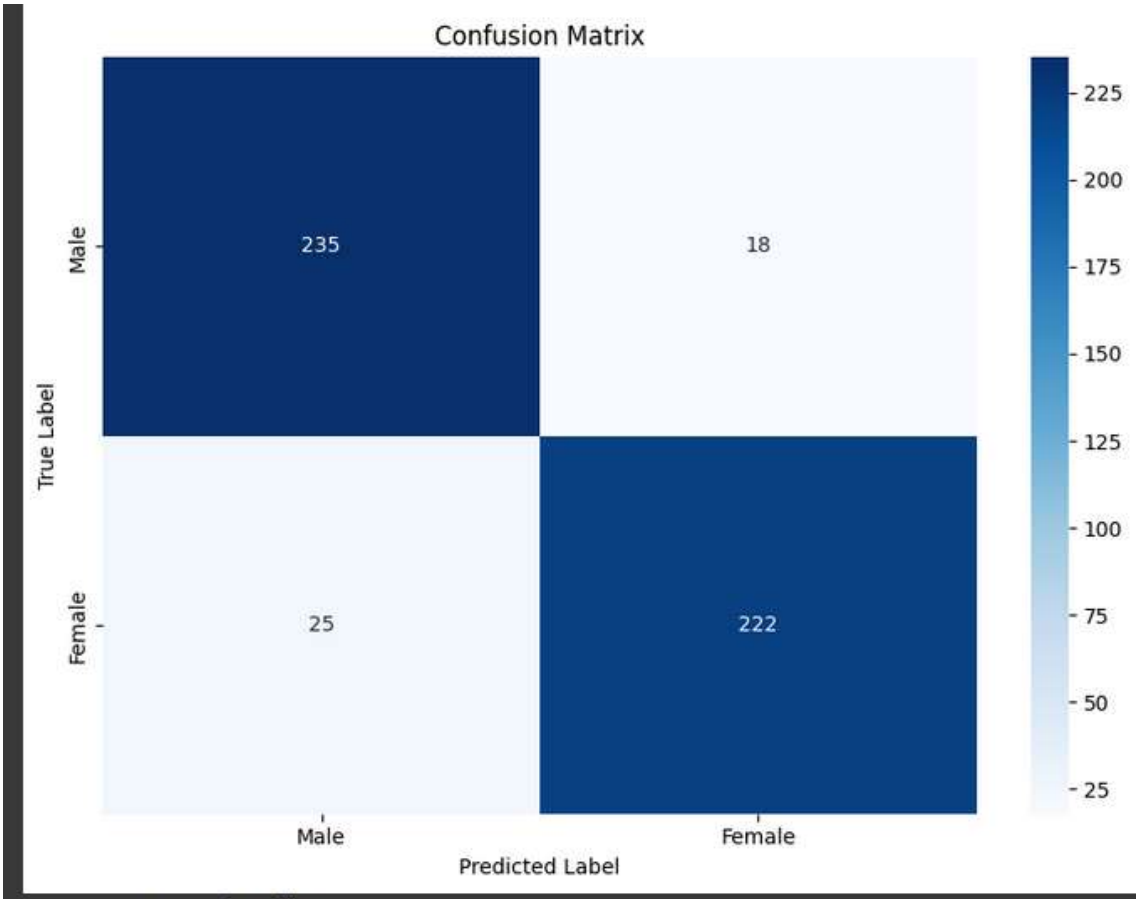
The transformer architecture is a simplified version of the standard Vision Transformer (ViT) encoder, tailored for image classification tasks. After the input image is divided into non-overlapping patches using a convolutional layer (patch_embed), each patch is linearly embedded into a vector of size embed_dim. Positional embeddings (pos_embed) are added to retain spatial information since transformers lack inherent positional awareness. These embedded patches, now treated as a sequence, are passed into a stack of Transformer encoder layers. Each encoder layer consists of multi-head self-attention mechanisms (with num_heads attention heads) and feedforward neural networks scaled by a multiplier (mlp_ratio) relative to the embedding size. This setup allows the model to learn long-range dependencies and relationships between different parts of the image. The use of batch_first=True ensures the input shape is (batch, sequence, embedding), making the transformer easily compatible with the preceding layers. Finally, the transformer outputs a contextualized representation for each patch, which is typically pooled (e.g., via mean pooling) for downstream classification.

TRAINING

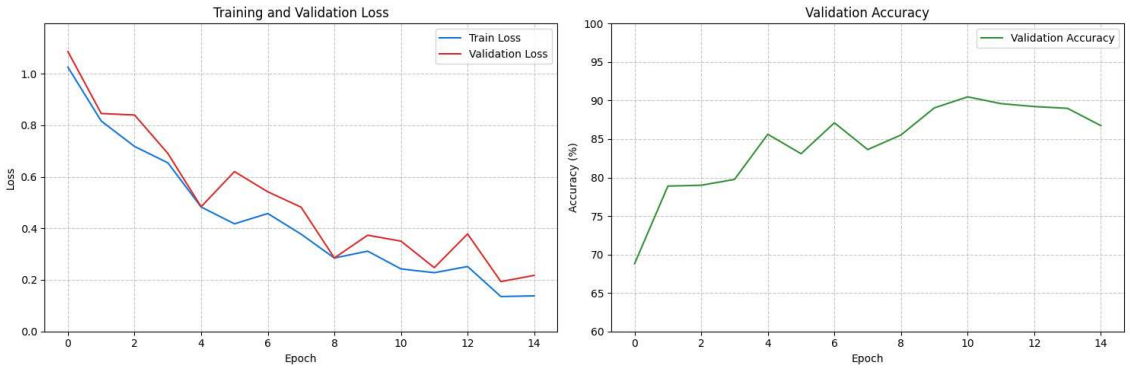
The model is trained over synthetic images. To train the model on a custom dataset a gui is included in the project along with scripts to evaluate accuracy and predict single instances. The training was done with a batch size of 32 for 30 epochs using a learning rate or 0.0001 and a weight deacy of 0.01. The project allows for custom values.

RESULTS

Evaluation:



Confusion Matrix



Accuracy curve

Prediction:

Male recognition

Female recognition

GITHUB LINK

https://github.com/jayabrotabanerjee/Simple-Face-Gender-Classifier