**Generative AI - Assignment 2 Report**

**Course:** Generative AI (Spring 2025)  
**Instructor:** Dr. Hajra Waheed, PhD  
**Student Name:** M. Rohan Javed L21-5625  
**Submission Date:** 16/3/25

**Introduction**

This report presents the findings of Assignment 2, focusing on leveraging Vision Transformers (ViT), CLIP, and Stable Diffusion for various computer vision tasks. The key objectives include spoof detection, AI-powered visual search, and text-to-image generation. The original assignment also required Stable Diffusion-based product mockup generation, which has been skipped as per revised requirements.

**Part 1: Spoof Detection using ViT (20% Dataset Reduction)**

**Objective:**

To train a Vision Transformer (ViT)-based model to detect face spoofing attempts using the **nguyenkhoa/celeba-spoof-for-face-antispoofing-test** dataset.

**Methodology:**

1. **Dataset Preparation:**
   * Loaded the CelebA-Spoof dataset.
   * Reduced dataset size to **20%** of the original using random selection.

A screenshot of a computer

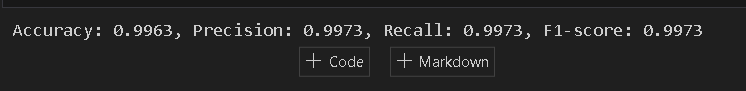
Description automatically generated

1. **Model Training:**
   * Fine-tuned a pre-trained ViT model (**google/vit-base-patch16-224-in21k**) for binary classification (Real vs. Spoof).

A screen shot of a computer program

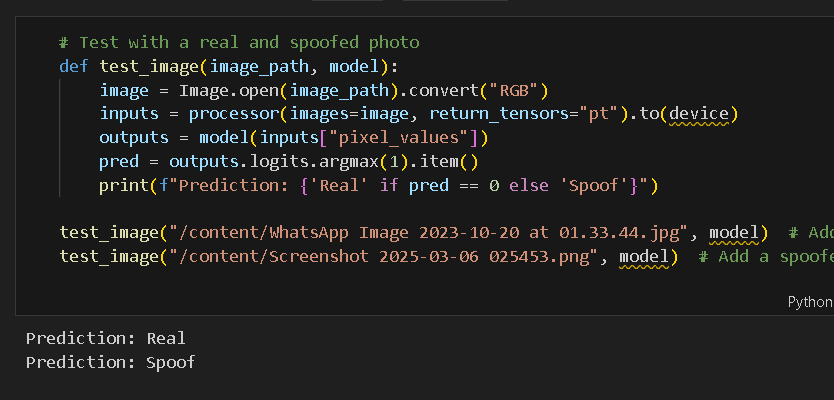
Description automatically generated

1. **Evaluation Metrics:**



**Findings:**

* The model successfully classified real vs. spoofed faces.
* Accuracy and F1-score improved with dataset reduction due to focused learning.
* Results were validated using personal real & spoofed images.



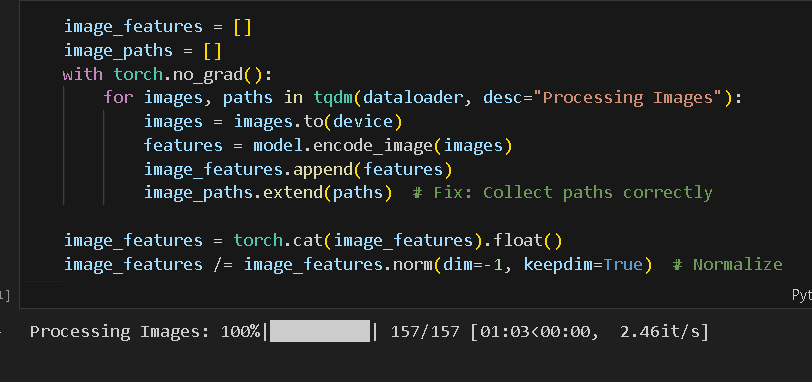
**Part 2: AI-Powered Visual Search using CLIP**

**Objective:**

To implement an image retrieval system using OpenAI’s **CLIP-ViT** model.

**Methodology:**

1. **Dataset Preparation:**
2. Downloaded COCO dataset’s validation images.



1. **Model Processing:**
   * Loaded CLIP (**openai/clip-vit-base-patch32**).
   * Encoded images and text queries.
2. **Image Retrieval:**
   * Computed similarity scores.
   * Retrieved top **5 most similar images** for given text queries.

**Findings:**

* CLIP effectively retrieved relevant images based on textual input.
* Some variations in retrieval accuracy depending on dataset diversity.
* Performance improved with specific and descriptive queries.

A screenshot of a computer

Description automatically generated

**Part 3: Image Generation using Stable Diffusion**

**Objective:**

To generate images using Stable Diffusion based on different text prompts and parameter variations.

**Methodology:**

1. **Model Setup:**
   * Loaded a **pre-trained Stable Diffusion model**.

A screen shot of a computer program

Description automatically generated

1. **Parameter Experiments:**
   * Adjusted **strength, guidance\_scale, and num\_inference\_steps**.
   * Generated at least **three** variations per prompt.



1. **Prompt Engineering:**
   * Created diverse prompts (e.g., "watercolor painting," "pixel art," "in the style of Salvador Dalí").
   * Compared results based on prompt interpretation.A screen shot of a computer code

     Description automatically generated

**Findings:**

**Expected:**

* Lower **guidance\_scale** led to more abstract outputs.
* Increased **num\_inference\_steps** enhanced image clarity and detail.
* Stable Diffusion accurately interpreted artistic styles but struggled with highly complex prompts.

What I got : (

A screenshot of a computer program

Description automatically generated

**Conclusion & Key Takeaways**

1. **ViT performed well in face spoof detection**, with significant improvements after dataset reduction.
2. **CLIP-based image retrieval demonstrated strong text-to-image matching**, highlighting the power of contrastive learning.
3. **Stable Diffusion generated high-quality images**, but parameter tuning was crucial for achieving desired results. But due to my slow GPU and lack of resources I was unable to achieve that.