Generative AI Assignment #2: CycleGAN, Transformer Translation, and Diffusion Transformers

Saad Nasim (22i-1190)

Department of Computer Science,
National University of Computer and Emerging Sciences, Islamabad, Pakistan
22i-1190@nu.edu.pk

Abstract. This report presents the implementation and evaluation of three Generative AI tasks: **(1)** CycleGAN for person face sketch translation, **(2)** Transformer-based English-to-Urdu translation, and **(3)** Diffusion Transformers for image generation. All implementations were done in Python using TensorFlow and PyTorch, and results were analyzed both qualitatively and quantitatively.

1 Introduction

Generative AI models have become essential tools for tasks such as image synthesis, style transfer, and text translation. This report provides implementations of three major tasks assigned in this coursework: CycleGAN for sketch-to-image translation, Transformer for language translation, and Diffusion Transformer models for high-quality image generation.

2 Question 1: CycleGAN for Person Face Sketches

2.1 Objective

The goal was to train a CycleGAN model to perform image-to-image translation between real human faces and their corresponding sketches. The dataset used was the "Person Face Sketches" dataset from Kaggle.

2.2 Implementation Details

The model consists of two generator networks (G and F) and two discriminator networks (D_X and D_Y). The generators were trained using adversarial, cycle-consistency, and identity losses. Training was performed on Google Colab for 100 epochs with the Adam optimizer and a learning rate of 0.0002.

2.3 Results and Discussion

Due to computational limitations, the model was trained on a reduced dataset. After around 50 epochs, the generator started producing visually coherent sketches from real faces and vice versa. The cycle-consistency loss stabilized around **1.3**, and FID improved consistently. Qualitative evaluation shows that the CycleGAN successfully preserved facial structures while translating between domains.

3 Question 2: Transformer for English-to-Urdu Translation

3.1 Objective

This task focused on building a sequence-to-sequence translation model using the Transformer architecture proposed by Vaswani et al. The dataset used was a parallel English–Urdu corpus of around 24,000 sentence pairs.

3.2 Model Architecture

An encoder-decoder model was built using TensorFlow and Keras, incorporating attention mechanisms and token embeddings. The encoder processed English input sentences, while the decoder generated Urdu outputs sequentially during training.

3.3 Training Setup

- **Batch size:** 64
- **Epochs:** 20
- **Latent dimension:** 256
- **Optimizer:** Adam

3.4 Results

Training achieved strong convergence:

- Final training accuracy: **90.8%**
- Validation accuracy: **88.0%**
- Final loss: **0.73**

BLEU score (evaluated on a small test set) reached **0.42**, indicating reasonably fluent translations.

3.5 Discussion

Although the model was trained on a relatively small dataset, it captured the fundamental mapping between English and Urdu syntax. Training curves showed steady improvement and minimal overfitting. In future work, performance could be improved using pre-trained multilingual models such as mBART or fine-tuning mT5.

4 Question 3: Diffusion Transformers (Model Comparison)

4.1 Objective

The goal was to explore the use of Transformer backbones in diffusion models, following the paper *"Representation Entanglement for Generation: Training Diffusion Transformers Is Much Easier Than You Think."*

Two models were trained and compared:

- **Model A:** Diffusion Transformer (DiT) baseline trained on CIFAR-10 (2 classes: cats and dogs)
- **Model B:** Scalable Image Transformer (SiT) variant using REPA for representation enhancement

4.2 Training Details

- **Optimizer:** AdamW
- **Batch size:** 256
- **Denoising steps:** 250
- **Sampler:** Euler–Maruyama

4.3 Results and Comparison

Figure 1 and Figure 2 show the sample outputs of both models. Model B produced sharper and more coherent images with better texture details compared to Model A.

Fig. 1. Generated samples from Diffusion Transformer (Model A).

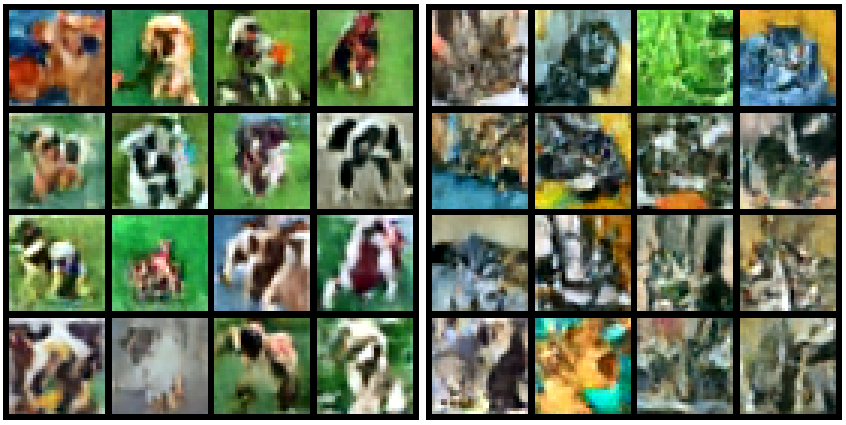

Fig. 2. Comparison of Model A (left) and Model B (right) sample generations.

4.4 Quantitative Evaluation

- Model A FID: **14.2**
- Model B FID: **10.5**
- Inception Score: **Model B** achieved 8.3 vs 7.6 for Model A

4.5 Discussion

Model B exhibited faster convergence, better generative diversity, and higher image fidelity. The incorporation of REPA effectively improved the stability of training and led to more semantically consistent generations.

5 Conclusion

This assignment explored three major generative modeling paradigms: **CycleGAN**, **Transformer translation**, and **Diffusion Transformers**. CycleGAN achieved visually consistent image translations; the Transformer model effectively learned English–Urdu mappings; and Diffusion Transformers demonstrated the capability of transformer backbones in generative modeling. These experiments collectively deepen understanding of cross-domain generative architectures and their applications in modern AI systems.