INFO8010: Anime Face Image generation

Ayoub Assaoud[1]

[1] *ayoub.assaoud@student.uliege.be (s207227)*

I. INTRODUCTION

For this project I propose to build a generative deep learning system that can generate anime faces. The system will be trained on a dataset of anime faces and will use hierarchical variational autoencoders as well as diffusion models to generate new faces. The goal is to create a system that can generate acceptable quality anime faces that are diverse and realistic by drawing samples, in the same time could be used for reverse problems like denoising by encoding-decoding the image. The project will also explore the use of different architectures and techniques to improve the quality of the generated faces.

II. OBJECTIVES

- drive a comparative study, to see which architecture is better for relatively small dataset like anime faces ($\leq 100K$ images).
- see the tradeoff between the reconstraction quality and the diversity of the generated faces (could use distangled $\beta - VAE$)
- Implement a hierarchical Variational Autoencoder (MHVAE) and a diffusion model for generating anime faces.
- Explore and compare some different architectures for the $p_\phi(z|x)$ and $p_\theta(x|z)$: *CNN, multi-head Transformer and MLP.*
- (Nice-to-have) implement a user interface to allow users to generate anime faces by sampling from the latent space or denoising.

III. DATA AND METHODOLOGY

A. Data

We will use the **Anime Face Dataset**, available at `https://www.kaggle.com/datasets/splcher/animefacedataset/`. This dataset has 63,632 ”high-quality” anime faces with total size of $415.18MB$.

B. Methodology

1. Data Processing

- Preprocessing anime face images (resizing, normalization).
- Augmenting the data to increase variability and robustness.
- **Possibly** additional images from another data set`https://www.kaggle.com/datasets/soumikrakshit/anime-faces` could be used for training ($21K$ images).

2. Model Architecture

We will explore and compare the following options:

- **PRIMARY Option**: Hierarchical Variational Autoencoder (MHVAE) - A hierarchical VAE with multiple latent variables sizes (slightely decreasing), where $p_\phi(z|x)$ and $p_\theta(x|z)$ are Vision Transformers (ViT as used in first project session) or simple CNNs, that predicts gaussian distribution parameters (μ, σ).
- **Option 2**: diffusion model using U-Net with ResNet blocks and self attention layer.
- **Option 3 (optional)**: if results are satisfactory and time is enough, a score-based model will be implemented.

3. Evaluation

- loss error is different for each model, for MHVAE we will use the ELBO loss, for diffusion model we will use the MSE loss
- the error between x and $\hat{x}$, is the mean squared error.

C. Infrastructure and Resources

Already have access to alan cluster.