Assignment 02: Stable Diffusion XL and Stable Diffusion XL Turbo
Due Date: 27Aug2024

Introduction

This assignment and paper are about the Stable Diffusion XL (SDXL) and Stable Diffusion XL Turbo (SDXLT) Machine Learning models. Both of these models are used to convert text input into a 2D color image. Both SDXL and SDXLT are based on the Stable Diffusion Model.

The Stable Diffusion (SD) model is an open-source text-to-image model released by Stability AI and has revolutionized the field of generative AI. The first major version was released in June 2022. The first generation of SD models have a resolution of 512x512 pixels and use a ViT-L/14 CLIP model for text training and have 860 million parameters.

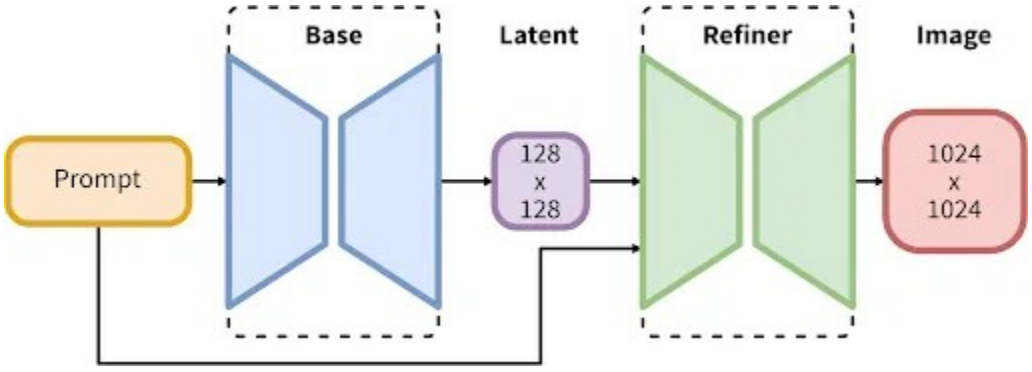
The 1st gen SD models are able to generate a wide range of styles and subjects and also have lower computational requirements. However they are not good at prompt comprehension and resolution. Subjects tend to be disfigured and are flat looking.

Over time the open-source SD model has been improved upon. We will discuss two such versions: Stable Diffusion XL 1.0 and Stable Diffusion XL Turbo

Stable Diffusion XL 1.0

This model was released in 2023 and provides outputs on the level of Midjourney and Dall-E and is able to be run on consumer level hardware. Images are a resolution of 1024x1024 pixels and rely on OpenCLIP-ViT/G and CLIP-ViT/L for text conditioning.

The model has a 3.5 billion parameter base model and a 6.6 Billion parameter ensemble pipeline.



Of all the Stable Diffusion Models, SDXL offers the highest resolution outputs and image quality. However the big downside is that this model requires powerful hardware to run locally.

Stable Diffusion XL Turbo

SDXLT is a lighter and faster version of SDXL. SDXLT sacrifices image quality in return for greater speed compared to SDXL. Like SDXL, SDXLT uses OpenCLIP-ViT/G and CLIP-ViT/L for

text conditioning, and has 3.5 billion parameters. However, SDXLT only takes one step to produce images compared to SDXL. SDXLT produces images with a resolution of 512x512 compared to SDXL which produces images 1024x1024.

Explanation of Code in Jupyter Notebook

Given 10 prompts, we process these prompts through the Stable Diffusion XL (SDXL) model and then through the Stable Diffusion XL Turbo (SDXLT) model. To measure performance, we compare the time it takes to process the 10 prompts to produce images from the same prompts.

We then perform an FID (Frechet Inception Distance) between the SDXL images produced and the SDXLT images produced. The SDXL images will be of a higher quality then the SDXLT so with FID, we are basically measuring the degradation from going from SDXL to SDXLT. The lower the FID score, the less degradation of image quality.

Results

Model	**Time to process prompts (sec)**	**FID Score**
SDXL	120-180	1.97
SDXLT	29	

As you can see, processing prompts through the SDXL prompt took over 4 times as long as processing images through SDXLT. The chart says that image processing took 2 minutes. This is more of an approximation. However it demonstrates that SDXLT is much more time efficient.

Examples of images produced are stored in 2 folders included with the code for this report as well as in a Jupyter Notebook.