Diffusion

Nidheesh Kannadasan/f20230310@goa.bits-pilani.ac.in

6th April 2025

1 Introduction

A diffusion model is most commonly used for generative tasks - images, audio, video, 3D models, etc. The model consists of the forward process called the diffusion process, where using Markov chain, Gaussian noise is gradually added to the data. The reverse process samples noise and iteratively removes noise to produce the output. An example is stable diffusion which uses a CNN based backbone called the U-Net. The U-Net is not scalabe because it keeps full spatial resolution features which leads to massive GPU usage and it struggles with limited receptive field and pipeline parallelism. The Transformer capitalizes on these weaknesses and hence the Diffusion-Transformer(DiT) was introduced.

2 Experiments on the official implementation

For the part a) of this task, the chosen label was 207(golden retriever) and CFG was set to 1(minimum), 4(Default), and 10(maximum) and the results respectively were

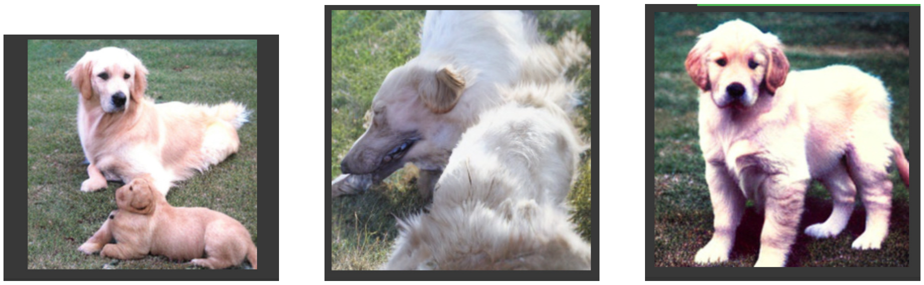

Figure 1: *
CFG = 1

Figure 2: *
CFG = 4

Figure 3: *
CFG = 10

Figure 4: Generated images of label 207 (golden retriever) at various CFG values

In Diffusion models, we not just want to generate random images but want to generate images with a prompt/text/class. Classifier free guidance(CFG) is a way to control how much influence does conditioning have on the output. Here class conditioning is used and higher the value of CFG, the more the generated image looks like the chosen class(golden retriever here), the results show the same with the rightmost image having the highest CFG resembling more like the golden retriever.

For part b) of this task, different number of sampling steps, 50,250 and 500 were tried on the class 207 with a CFG of 10 and here are the results respectively:

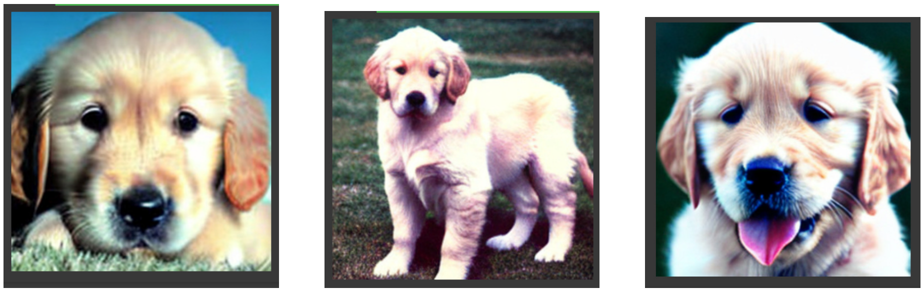

Figure 5: *
50 Sampling steps

Figure 6: *
250 Sampling steps

Figure 7: *
500 Sampling steps

Figure 8: Effect of varying sample count on class 207 generation with CFG = 10

The number of sampling steps affects the time it takes to generate the images and the quality of generation. The higher the sampling steps, the longer it takes to generate the image and hence the best quality as can be seen in the rightmost image.

3 Note

I attempted task 2 and changed the attention block to xformers but it required higher computational power than colab's free T4 GPU and my device doesnt have a GPU and henceforth i wasn't able to proceed with the following tasks. I think it is important to understand the concept than to write code and so i have read upon the metrics in the following tasks such as CLIP, CMMD, SigLIP,etc