1 Image Synthesis with Diffusion Model

Please follow the instructions in this notebook. You will implement a text conditioned denoising unet and the sampling algorithm. Then you'll see the diffusion process on a single image and try to train your own diffusion mode. Once you finished with the notebook,

- download submission.zip and submit it to Gradescope.
- Answer the following questions in your submission of the written assignment:

(a) Why might a U-Net by a good choice for the model backbone? List two reasons.

Solution: (This is an open question, if the explanation of the answer is reasonable, then it can be considered correct)

- The model must output the amount of noise in the input, which is of the same shape as the input, meaning both the input and output are of identical shape. This is exactly what a U-Net does.
- UNet has an inductive bias geared towards image-like data. This is due to using convolutions and the downsampling followed by upsampling with skip connections between them which gives it the ability to pick out hierarchical and spatial structures in data.

(b) Screenshot your visualization for get noisy image and include it in your submission of the written assignment. And describe the picture you observed briefly. What kind of process is this?

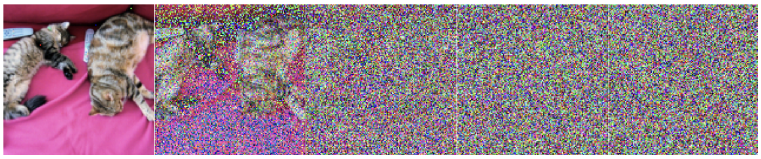

Abbildung 1: Visualization for Sampling

Solution: Shown as Abbildung 1. This should be a process of gradually adding noise to the original picture. Corresponding to the q sample in diffusion modelmodel algorithm.

(c) Screenshot one of your visualization about the sampled images and include it in your submission of the written assignment. Then answer the following question. How does the model perform and does it meet your expectations? If not, what do you think are the directions for improvement?

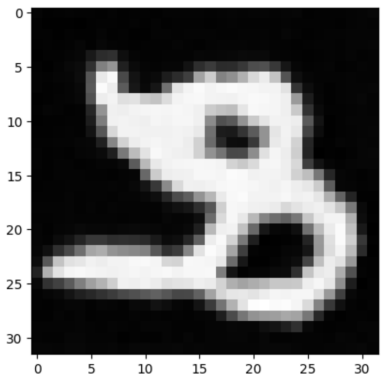

Abbildung 2: Visualization for Model Performance

Solution: The performance of the model was not so great and did not meet expectations. As can be seen from the visualization picture, it only learns the style of the dataset, but does not learn the real semantic information. The main reason is that the datasets and models we use are too small. At the same time, MNIST is not a standard image-text pair dataset, so our text definition is not very accurate. In addition,

clip does not perform particularly well with handwritten digits. In order to improve the performance of the model, we also need to consider using a larger and more professional image-text pair dataset while increasing the scale of the model. This further underscores the current trend towards larger models.