Fine-tuning Diffusion Models

Marena Keys
Advisor: Dr. Xiaoming Huo

Georgia Tech College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

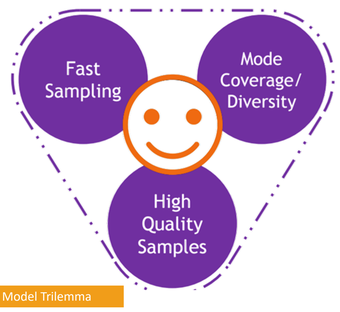
Github

Objective

- Generative AI models are increasingly growing in popularity due to its powerful ability to learn complex datasets.
- The aim of this project is to learn how to fine-tune a diffusion models and understand the architecture of diffusion models.
- Main focus is text-to-image models

Background: Diffusion Models

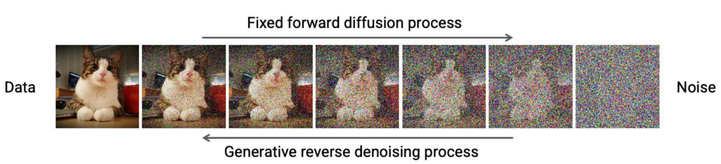
- Diffusion models are generative models that produce new data similar to its training data
- Stable diffusion is a diffusion model that generates new images from text prompts



Diffusion Model Trilemma

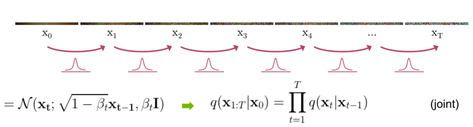
Denoising Diffusion Probabilistic Models

- DDPMs work by gradually adding noise (pixels) to the input (images) and then learning to denoise in order to generate new images
- The process can be defined as a Markov chain. The next image that is created by adding noise or denoising only depends on the previous image
- *The image below shows forward diffusion and reverse diffusion*



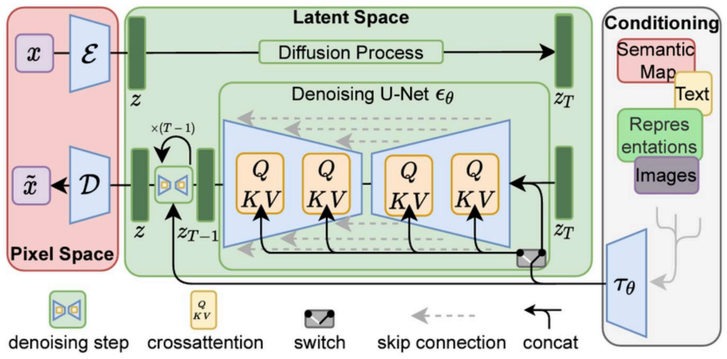
Forward Diffusion

At each step, a Gaussian (Normal) distribution is used to generate the next noisy image based on the previous image.



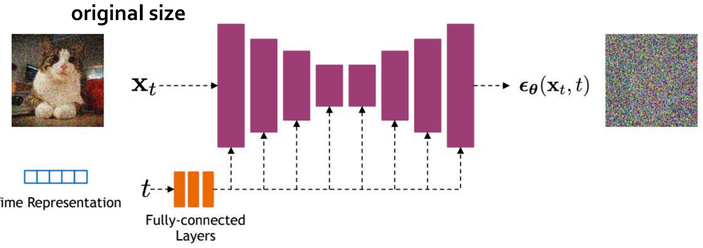
Latent Diffusion

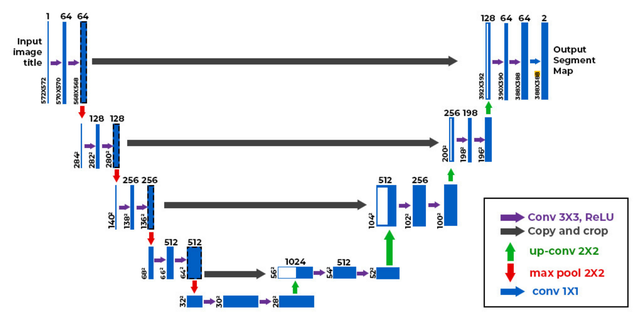
- Designed to show the internal structure of a dataset
- Maps high-resolution data to a low-dimensional latent space



U-Net Architecture

- Symmetric (U-shape) structure consisting of four encoders and four decoders
- The input (image) is initially downsampled as it goes through each encoder. The decoder upsamples the image until it's back to the original size

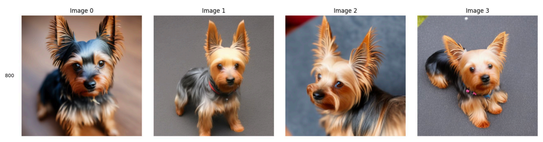




Fine-tuning: DreamBooth

- DreamBooth is a machine learning technique where one can train (fine-tune) a pretrained model to create a new set of images
- The images below were generated using Google Colab

From Colab: "A photo of cat inside a kitchen, with food, hyper-realistic, hyper-detailed"



From Colab: "photo of yorkie outside, sunny day, clear, ultra photoreal, insanely detailed, 8k, crisp, brilliant

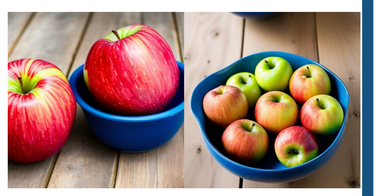
Results

- From the COCO dataset, a smaller dataset consisting of one hundred images along with their captions was created. The captions are stored in a metadata.jsonl file
- A fine-tuned model was constructed using the custom dataset in order to generate new images with captions

Original

"A large white bowl of many green apples."

New



"Apples in a blue bowl sitting on a brown counter."

Sources

1. *Common Objects in Context*. COCO. (n.d.). https://cocodataset.org/#download
2. Lin, T.-Y. (2017). *Cocoapi/pythonapi/pycocodemo.ipynb at master · cocodataset/cocoapi*. GitHub. https://github.com/cocodataset/cocoapi/blob/master/PythonAPI/pycocoDemo.ipynb
3. Ronneberger, O., Fischer, P., & Brox, T. (2015, May 18). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv.org. https://arxiv.org/abs/1505.04597
4. Vahdat, A., Kreis, K., & Gao, R. (2022, June 19). *Denoising diffusion-based generative modeling: Foundations and Applications*. Denoising Diffusion-based Generative Modeling: Foundations and Applications Tutorial. https://cvpr2022-tutorial-diffusion-models.github.io/
5. What is stable diffusion? - stable diffusion AI explained - AWS. (n.d.). https://aws.amazon.com/what-is/stable-diffusion/