Summary of Stable Diffusion

Ari Singer (arising), Vishwa Ramanakumar (vishram), Jack Holland (jackholl)

Problem and Motivation

Introduced in 2015, Diffusion Models (DMs) are a type of generative model that attempt to recreate original information from noisy data by reversing the noising process. While such models have been popularly associated with a growing trend of AI artwork, they also have had other groundbreaking applications for things such as 3D molecule generation. However, one of the primary drawbacks of DMs are that they are computationally expensive given the need for repetitive function evaluation and gradient computations as well as the high dimensional nature of RGB image data. To combat this and enable the training of DMs on limited computational resources while minimally sacrificing model performance, the introduction of a latent space is proposed, creating a Latent Diffusion Model (LDM).

Related Works

UNet

Developed in 2015, the UNet is a convolutional neural network used for biomedical image segmentation. In general, it is useful in predicting image noise. Within the context of this problem, if a noisy image is supplied to the UNet along with the level of noise (magnitude of perturbation), the noise added to the original image is predicted. The model can be divided into two sections: a contracting encoder layer and an upsampling decoder layer. (Fig. 1)

In the encoder, max-pooling is utilized for convolution. The decoder takes the output of the final encoding layer and uses the convoluted version of each image in the corresponding encoding layer to output noise images of higher resolution. The UNet network is the backbone of the Latent Diffusion Model.

CLIP

CLIP, standing for Contrastive Language-Image Pre-Training, is an open-source, multimodal model that trains on pairs of images and text with the ultimate goal of matching an image with the most relevant text or vice-versa. As the name suggests, it is contrastive in nature, meaning that more similar text-image pairs are closer together in the latent space.

There are two major components to the CLIP model: the text encoder and the image encoder. During training, the text encoder, which is a transformer, takes in a series of texts and outputs an embedded vector. Similarly, the image encoder, which can be either a ResNet network or Vision Transformer model, takes a series of images and outputs an embedded vector. (Fig. 2)

The cosine similarity score of each pair of text and image embeddings is calculated. The objective of the model is to maximize the similarity score of the matching image-text pairs, or the diagonal values, while minimizing the score of the other pairs. Two separate cross entropy loss values are used to represent these values, and the final loss is the average of both. This is a symmetric loss function.

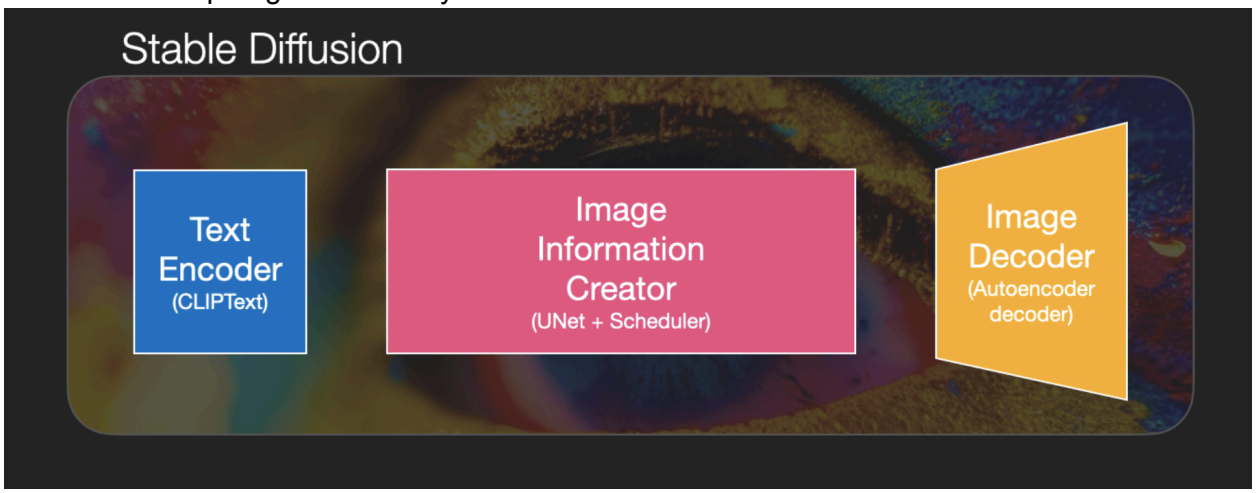
To predict using CLIP, an image is fed in along with a series of texts, and the text with the highest similarity score is returned.

Solution Overview

Diffusion is the net movement of anything (for example, atoms, ions, molecules, energy) generally from a region of higher concentration to a region of lower concentration. In this context, diffusion can be thought of as an object becoming noisier. (Fig. 3)

The goal of Stable Diffusion is to reverse this process, iteratively removing noise such that the resulting object is noise free. The ability to *predict* the noise component of an object allows us to remove it. (Fig. 4)

Stable Diffusion consists of a pipeline of components which intend to produce an image based on the textual input[1] given to the system.



The “Text Encoder” component, CLIPText, is what creates a vector embedding of the textual information given to the system. This embedding is used within the “Image Information Creator” component to push it towards relevant output.

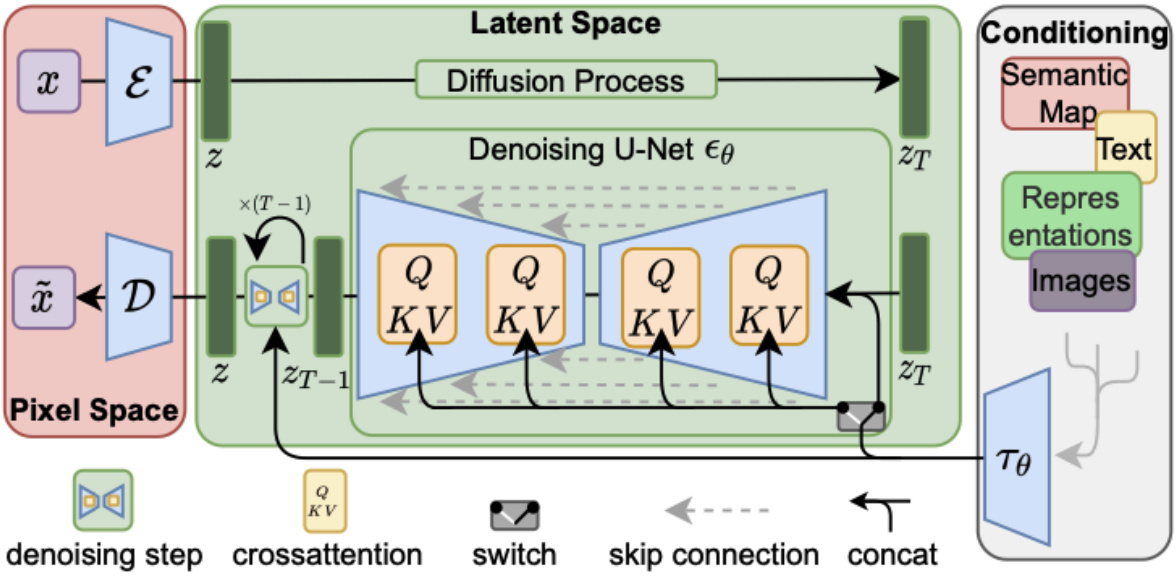
The middle component, the “Image Information Creator”, or UNet, is where the performance gain over previous models is achieved. It is in this component where noise is predicted and iteratively removed, and it acts entirely within the latent space, which is a much smaller dimensionality than an RGB image. *The usage of a latent space is a major innovation employed by Stable Diffusion*, as it, compared to the high-dimensional pixel space, is more suitable for likelihood-based generative models, as they can now (i) focus on the important, semantic bits of the data and (ii) train in a lower dimensional, computationally much more efficient space.

Finally, the “Image Decoder” component takes the final latent space representation from the “Image Information Creator” component and upscales the resolution into a final RGB Image.

Training Stable Diffusion is a multi step process. Firstly, an autoencoder decoder model is trained to provide good encode/decode functions to and from the pixel space and latent space. Once this is achieved, the UNet must be trained to effectively predict noise. A train image, x, is encoded into the latent space (as z), then diffused by overlaying random noise (as z_T). Gradient descent is then used to tune the UNet upon successive iterations. Notice that the conditioning

[1] This application, Text-to-image (textual conditioning), is just one of the ways this work can be employed.

step, τ_θ, is baked into the UNet, and need not be limited to just textual conditioning.



Once training is complete, the model can be used to generate images by simply sampling random noise within the latent space, and plugging it in as z_T, and passing it through the bottom pipeline.

Limitations
Through the incorporation of a text encoder such as CLIP, stable diffusion also inherits its drawbacks including its computation-intensiveness. As such stable diffusion will struggle with complex/abstract/systematic prompts. For example a stable diffusion model will have difficulty when tasked to have some number of items in a picture as could be seen in a prompt to a stable diffusion model to generate “five dogs in a pool” where only three were generated. (Fig. 5)

Stable diffusion models are also found to have bad performance for some small details, such as the well known and documented inability to correctly create peoples’ fingers.

Future Research Directions
As stable diffusion models can be rather small (~4GB - 12GB), it is entirely feasible to run models locally, unlike what is seen for large language models (GPT-3) that are usually seen as hosted services. There is potential research to bridge this divide.

The overall research space of stable diffusion is also lacking in relation to that of LLMs and there are many research directions such as that stable diffusion to generate different types of output data (some current research involves music/video generation).

Further works into CLIP can also be utilized to improve diffusion models including topics such as object detection, image search, image segmentation, and more.

Summary of Class Discussion
It was brought to question whether it would be conceivable to generate text through stable generation. It was noted that diffusion lacks the length flexibility of LLMs due to the existing length restriction; however this is only a hard limit that could allow for text generations of length

up to that limit. A few attempts are currently in process but it is entirely likely that the potential solutions have not been exhausted yet.

Another question of particular note was if there is a way to cache some intermediate noise steps to save compute during diffusion inference. In response to this, it was pointed out that multiple papers have been released that utilize some caching scheme to optimize the stable diffusion model.

It was also questioned if other portions of the model could be trained, frozen, and then plugged into any diffusion model such as how CLIP is being utilized in the model as described.

Appendix

Fig. 1

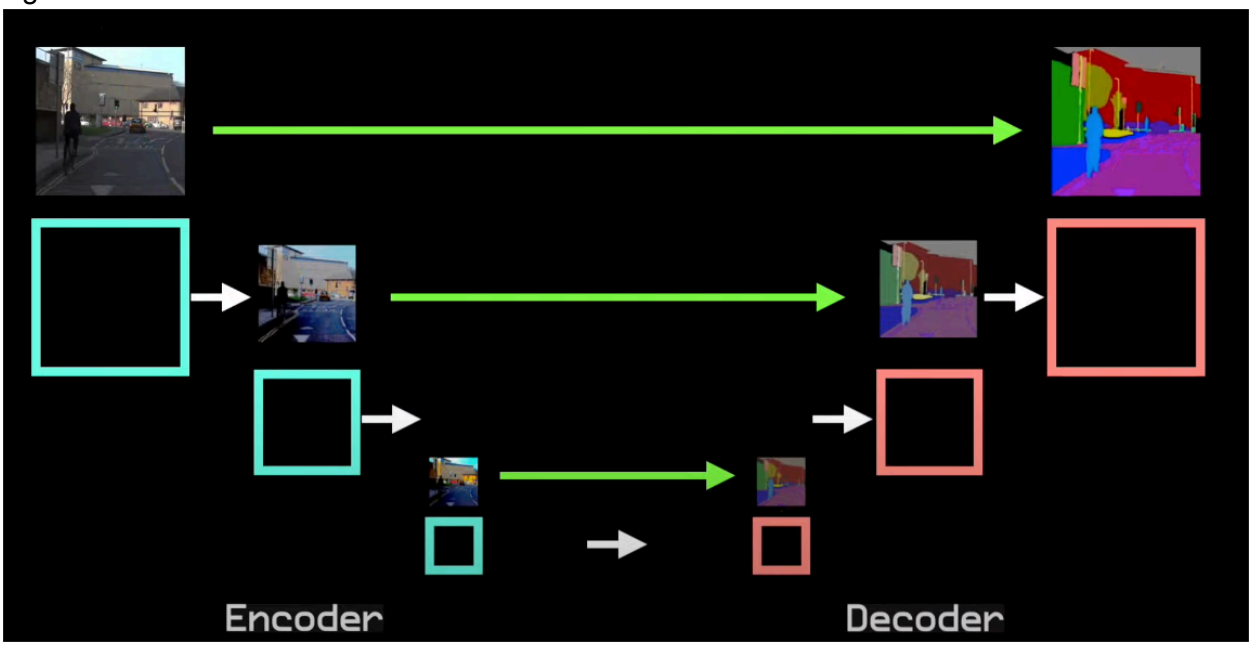


Fig. 2

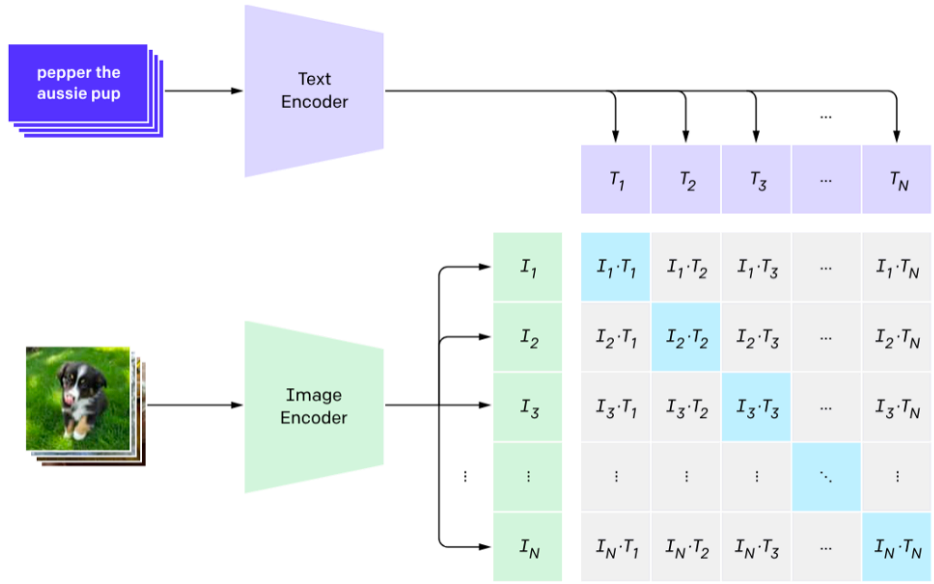


Fig. 3

Fig. 4

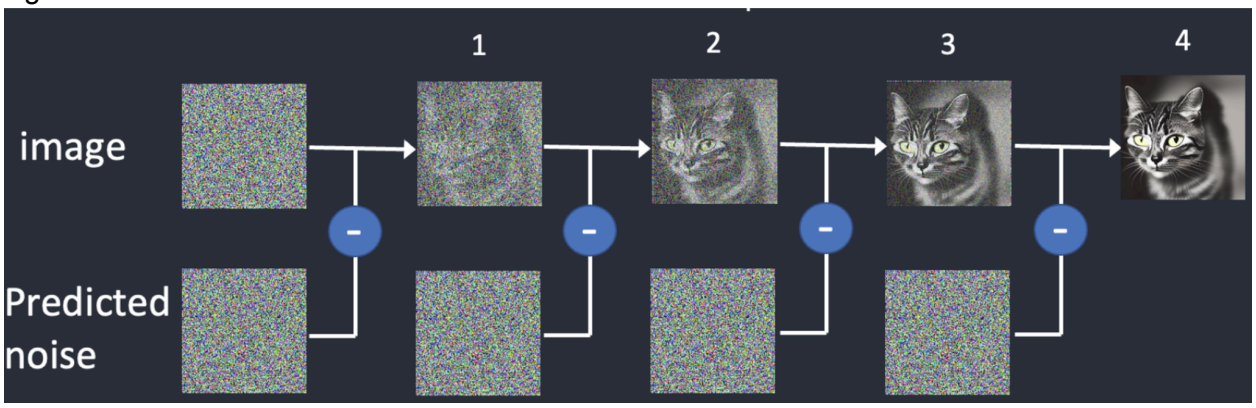


Fig 5.