STABLE DIFFUSION TEXT-IMAGE GENERATION

Syed sha alam.A[1], Jeyamurugan.N[2], Mohamed faiz ali.B[3],Veerasundari.R[4]

[1,2,3,4]Veltech HighTech Dr.Rangarajan Dr.Sakunthala Engineering College

Abstract— A text-to-image technology called Stable Diffusion will enable billions of individuals to produce beautiful works of art in in a few seconds. It can run on GPUs available in consumer gadgets and is an improvement in both performance and quality. It will advance democratic. image generation by enabling both researchers and the general public to use it under various conditions. We anticipate the open ecosystem that will grow up around it as well as new models to really probe the limits of latent space.

Keywords*— laion5B, Perceptual Image Compression,Latent Diffusion Models, Conditioning Mechanisms , Text to image generation, Image Modification*

INTRODUCTION

In 2022, the Stable Diffusion deep learning text-to-image model was made public. Its major function is to produce comprehensive images based on text descriptions, while it may also be used for other tasks including inpainting, outpainting, and making image-to-image conversions prompted by a text prompt.In conjunction with EleutherAI, LAION, and StabilityAI, a text-to-image, text-to-video model called Stable Diffusion was created to convert natural language descriptions into digital images. The technique can also be applied to other tasks, such as producing image-to-image translations that are instructed by text. It was praised by PC World as "the next killer programme for your PC" and is compatible with the majority of consumer hardware with a basic GPU. It received training using 512*512 the LAION-5B database's images. This model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. Additionally, it spent $600,000 on training using 256 Nvidia A100 GPUs.

LITERATURE SURVEY

Diffusion models (DMs) achieve cutting-edge synthesis results on image data and beyond by breaking down the creation of images into a series of denoising autoencoder applications. Their approach also makes it possible to control the image-generating process without retraining using a guiding mechanism. However, as these models usually operate in pixel space, effective DM optimization can require hundreds of GPU days, and inference is expensive due to sequential evaluations[3]. In order to facilitate DM training on limited computing resources while keeping their quality and flexibility, we employ them in the latent space of powerful pretrained autoencoders. In contrast to previous research, training diffusion models on such a representation allows for the first time to achieve a close to ideal point between complexity reduction and detail preservation, significantly improving visual perception. We transform humans into artists by incorporating cross-attention layers into the model design. Because it made a high-performance model accessible to the general public (performance in terms of image quality, as well as speed and relatively low resource/memory requirements), [3]

Currently, Stable Diffusion is disturbing artists all over the world and motivating the open source machine learning community[8]. I was interested to discover what other applications this significant technological advancement might have besides making working designers and artists worry about their future employment. I discovered that the model makes for a very potent lossy picture compression codec while exploring with it. Here are some findings compared to JPG and WebP at high compression factors, all at a resolution of 512x512 pixels, before outlining my methodology and sharing some code: These illustrations clearly show that, as compared to JPG and WebP, Stable Diffusion compression yields photos with a much greater level of image quality at a lesser file size. There are some significant limitations to this quality.[8]

Diffusion models can be thought of as low-dimensional learning manifolds, generally images, that map points in a high-dimensional latent space. The noise scheduling strategy used during pre-training can be used to interpret the middle values between the picture manifold and the latent space as noisy images[3]. This interpretation allows us to introduce Boomerang, a local image manifold sampling technique that makes use of diffusion model dynamics. The procedure is referred described as a "boomerang" because an input picture is first fed noise to get it closer to the latent space before being brought back to the image space via diffusion dynamics. Using this method, we generate pictures that, although not exact duplicates, are equivalent to the original input images on the image manifold. How much noise we choose to add will affect how similar the resulting image is to the original. Additionally, the generated pictures' stochasticity allows us to We can repeat any number of local samples without repeating any data. With these phrases, we show Boomerang in three different contexts. First, we present a system for building datasets with variable degrees of anonymity and privacy protection. Then we show how to use Boomerang for data augmentation while still staying on the picture manifold.The framework for image super-resolution with 8x upsampling is introduced in the third section. Boomerang can be utilised with pretrained models on a single, affordable GPU and does not call for any changes to the training of diffusion models.[3]

MATERIALS AND METHODS

Diffusion Models

Since diffusion models are generative models, they produce data that is similar to the data used to train them. Fundamentally, the way that diffusion models work is by corrupting training data by gradually adding Gaussian noise, and then learning to recover the data by undoing this noise process.

Latent Diffusion

Diffusion models that are trained in a latent space are known as latent diffusion models. A latent space is a multidimensional abstract area that contains a valuable internal representation of events that have been observed from the outside. Samples that are equivalent in the real world are placed next to one another in the latent space. Its main goal is to transform raw data, like the pixel values of a picture, into an acceptable internal representation or feature vector so that the learning subsystem, frequently a classifier, can identify or categorise patterns in the input.

PERCEPTUAL IMAGE COMPRESSION

It can be difficult to create an efficient perceptual image compression algorithm. When developing such an algorithm, it is crucial to take into account that people are able to see a considerable level of detail in photos. The resolution of the image, the colour scheme, and the distribution of pixel values are some of the aspects that must be considered while constructing a perceptual image compression algorithm. Based on earlier research, our perceptual compression model employs an auto encoder trained by combining a perceptual loss with a patch-based adversarial objective. This prevents blurriness from being introduced by relying solely on pixel-space losses and ensures that the re-constructions are contained to the image manifold by enforcing local realism L2 or L1 objectives, for example. More specifically, the encoder E encodes an image x that is RHW3 in RGB space into a latent representation z = E(x), and the decoder D reconstructs the picture from the latent, producing x that is equal to D(z) = D(E(x)), where z is RhWc. We study various down sampling factors f = 2m, with m N, because the encoder down samples the image by a factor f = H/h = W/w. We test two types of regularizations in order to avoid latent spaces with arbitrarily high variance. As with a VAE, the first form, KL-reg., applies a minimal KL-penalty to the learned latent whereas VQ-reg. makes use of a vector quantization layer within the decoder. Although this mod-el with the decoder absorbing the quantization layer. We may apply quite moderate compression rates and produce extremely good reconstructions since our future DM is built to operate with the two-dimensional structure of our learned latent space, z = E(x). This is in contrast to earlier efforts [23, 64], which neglected much of the intrinsic structure of z by using an arbitrary 1D ordering of the learned space z to predict its distribution auto-regressively. As a result, our compression model better preserves x's details. The supplement contains all of the information about the goal

CONDITIONING MECHANISMS

Diffusion models are theoretically capable of modelling conditional distributions of the form p(z|y), just like other kinds of generative models. The synthesis process can be controlled by inputs y like text, semantic maps, or other image-to-image translation tasks. In the context of image synthesis, however, combining the generative power of DMs with other types of conditionings beyond class-labels or blurred variants of the input image is so far an under-researched area of study.

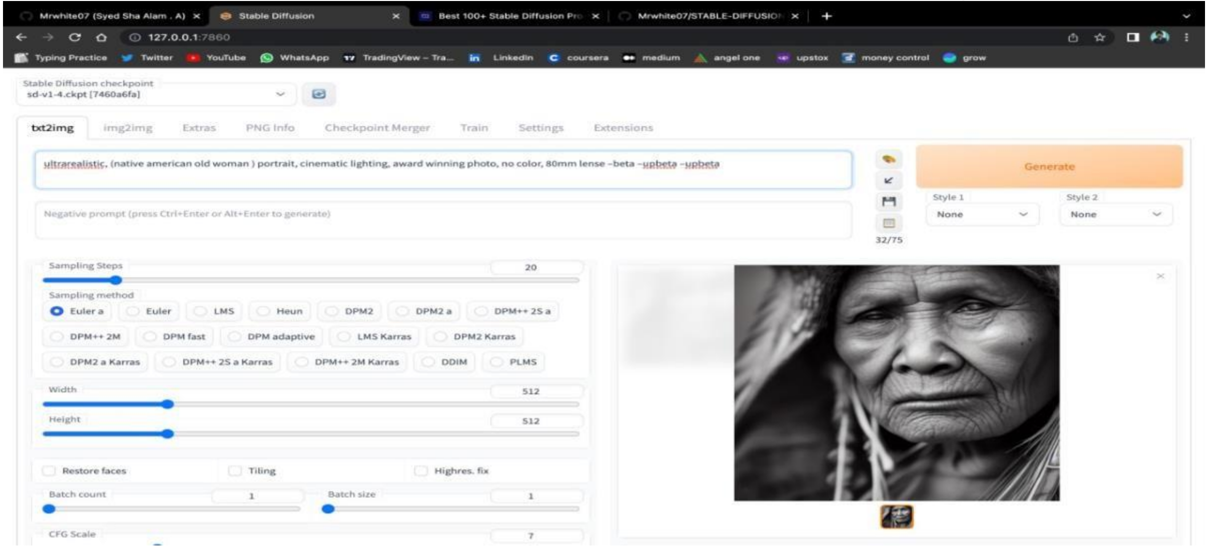
By adding the cross-attention mechanism to the basic UNet backbone of DMs, which is useful for learning attention-based models of diverse input modalities, we make DMs into more adaptable conditional picture generators. In order to pre-process y from many modalities (like language cues), y is projected by a domain-specific encoder to an intermediate representation (y) RM d, which is then mapped to the intermediate layers of the UNet by implementing a cross-attention layer. Attention(Q,K,V)=softmaxQKT V, with a d Q is equal to W(i) + i(zt), K is W(i) + i(y), and V is W(i) + I (y). QKV In this case, WV Rddi, W I Rdd, and W I Rdd are learnable projection matrices. I (zt) RN di signifies a (flattened) intermediate I representation of the UNet implementation. to provide a picture. samples with a resolution of 256 256 from the LDM strain edonCeleb AHQ, FFHQ, LSUN-Churches, LSUN-Bedrooms, and class conditional ImageNet. Zoomed-in viewing is recommended. See the Appendix for additional samples. The conditional LDM is then learned using LLDM based on image-conditioning pairs: = EE (x), y, N (0, 1), t. (3)

where are both present combined optimization using Eq. Since may be parameterized with domain-specific experts, such as (unmasked) trans- formers when y are text prompts, this conditioning process is adaptable.

.

TEXT TO IMAGE GENERATION

The Stable Diffusion "txt2img" text to image sampling script takes a text prompt as well as many option parameters for sample kinds, output picture dimensions, and seed values. In accordance with the model's interpretation of the prompt, the script generates an image file. Invisible digital watermarks are applied to created photos to help users recognise them as being the work of Stable Diffusion, however these watermarks lose their effectiveness when an image is cropped or rotated**.**



Stable Diffusion checkpoint: sd-v1-4.ckpt [7460a6fa]

ultrarealistic, (native american old woman) portrait, cinematic lighting, award winning photo, no color, 80mm lense –beta –upbeta –upbeta

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Style 1: None
Style 2: None

Sampling Steps: 20

Sampling method:
● Euler a ○ Euler ○ LMS ○ Heun ○ DPM2 ○ DPM2 a ○ DPM++ 2S a
○ DPM++ 2M ○ DPM fast ○ DPM adaptive ○ LMS Karras ○ DPM2 Karras
○ DPM2 a Karras ○ DPM++ 2S a Karras ○ DPM++ 2M Karras ○ DDIM ○ PLMS

Width: 512
Height: 512

□ Restore faces □ Tiling □ Highres. fix

Batch count: 1
Batch size: 1

CFG Scale: 7

Fig 1:text2img

Each txt2img generation will use a unique seed value that will have an impact on the final image. Users can choose to use a different seed to experiment with various created outputs or stick with the same seed to get the same image output as a prior image generation. The number of inference steps for the sampler can also be changed by the user; a larger value takes more time, but a smaller value could lead to visual flaws. The user can also control how closely the output image resembles the prompt by adjusting the classifier-free guidance scale value. [20] While use cases aiming for more specified outputs may choose to use a higher number, more exploratory use cases may choose to use a lower scale value.

More text2img features are offered by front-end Stable Diffusion implementations that let users change the weights assigned to different areas of the text prompt. By surrounding keywords in brackets, emphasis markers allow users to increase or decrease the emphasis on a keyword. "Negative prompts" are another way to change the weight given to different portions of the prompt. Some front-end implementations allow users to define suggestions that the model should avoid while creating images using a feature called negative prompts. The specified prompts could be undesirable image elements that would otherwise be present in image outputs as a result of the user's or the model's initial design's positive prompts.

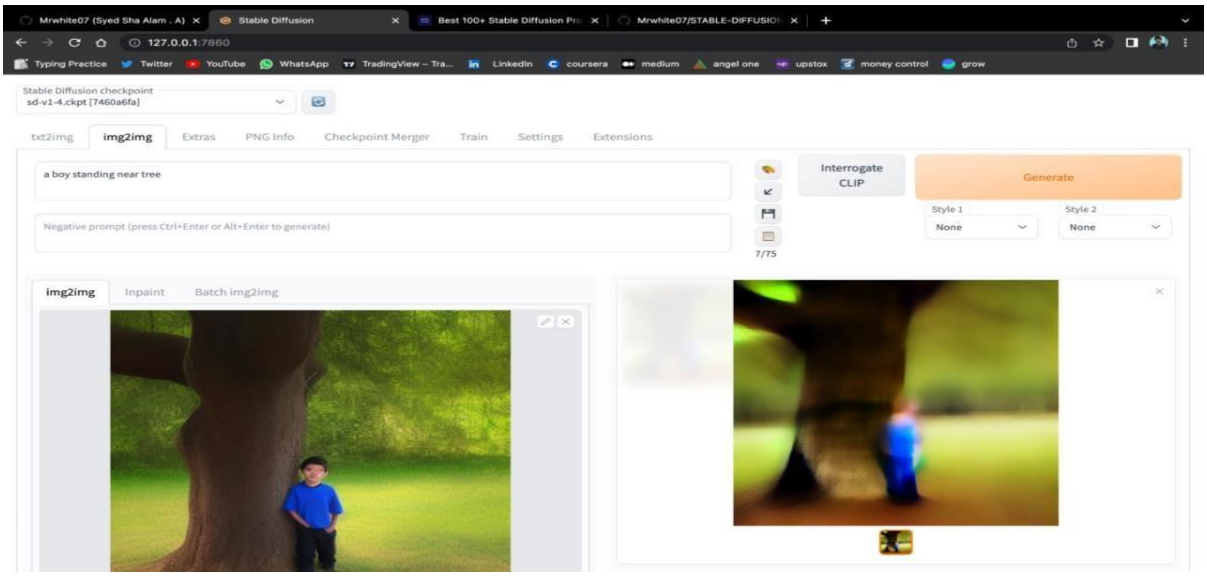
FUTURE ENHANCEMENT

IMAGE TO IMAGE:

The concept behind text-to-image production also applies to image-to-image artificial intelligence art (img2img). For the AI, users still type in prompts. The primary distinction between these two is that a base image is used in the production of an img2img file.

Click the 'img2img' tab at the top to utilise img2img. Drag and drop the image into the "Image for img2img" column in the left column or click it to pick it to load it. Stable Diffusion will then create a new image based on the loaded image after you enter text in the prompt and click "Generate."

•



Stable Diffusion checkpoint: sd-v1-4.ckpt [7460a6fa]

a boy standing near tree

Negative prompt (press Ctrl+Enter or Alt+Enter to generate)

Style 1: None

Style 2: None

Fig2: img2img

ARCHITECTURE DIAGRAM

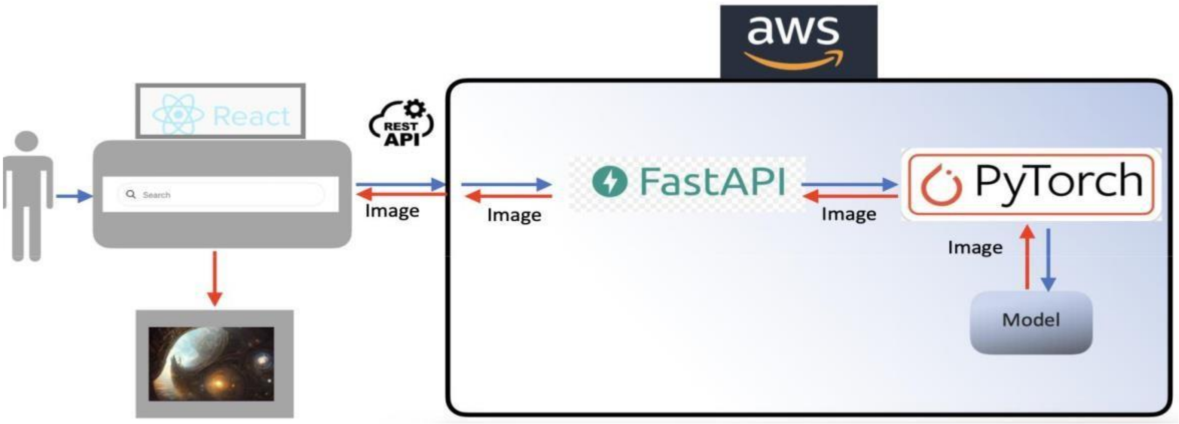


Fig3:Architecture of stable diffusion text to image

A variation of the diffusion model (DM) known as the latent diffusion model is used by stable diffusion (LDM). Diffusion models, which were first used in 2015, are trained with the goal of eradicating many applications of Gaussian noise on training images, which can be compared to a series of denoising autoencoders. The variational autoencoder (VAE), U-Net, and an optional text encoder make up Stable Diffusion. The VAE encoder condenses the image from pixel space to a more basic semantic meaning in a less dimensional latent space. Forward diffusion involves repeatedly applying Gaussian noise to the compressed latent representation.

The output of forward diffusion backwards is denoised by the U-Net block, which is made up of a ResNet backbone, in order to obtain latent representation. The VAE decoder then converts the representation to produce the final image. return to the pixel world. Flexible conditions can be applied to the denoising step based on a list of text, a picture, and other modalities. Through a cross-attention mechanism, the en-coded conditioning data is made available to denoising U-Nets. The fixed, pretrained CLIP ViT-L/14 text encoder converts text prompts to an embedding space for text conditioning. As a benefit of LDMs, researchers point to increased computational efficiency for training and generation.

.

CONCLUSION

Latent diffusion models are a quick and easy technique to boost the training and sampling effectiveness of de-noising diffusion models without sacrificing their quality. Our experiments could demonstrate favourable results compared to state-of-the-art methods across a wide range of conditional image synthesis tasks without task-specific architectures based on this and our cross-attention conditioning mechanism.

REFERENCES

1. "Stable Diffusion Repository on GitHub". CompVis - Machine Vision and Learning Research Group, LMU Munich. 17 September 2022. Re- trieved 17 September 2022.

2. RunwayML. "stable-diffusion-v1-5". *Hugging Face*.
3. "Diffuse The Rest - a Hugging Face Space by huggingface". *hugging-face.co*. Archived from the original on 2022-09-05. Retrieved 2022-09-05.

   Rombach, Blattmann; Lorenz; Esser; Ommer (June 2022). *High-Reso-lution Image Synthesis with Latent Diffusion Models* (PDF). International Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA. pp. 10684–10695. arXiv:2112.10752.

4. "Stable Diffusion Launch Announcement". *Stability.Ai*. Archived from the original on 2022-09-05. Retrieved 2022-09-06.

5. "Revolutionizing image generation by AI: Turning text into images". *LMU Munich*. Retrieved 17 September 2022.

6. Wiggers, Kyle. "Stability AI, the startup behind Stable Diffusion, raises $101M". *Techcrunch*. Retrieved 2022-10-17.

7. *Stable Diffusion*, CompVis - Machine Vision and Learning LMU Mu-nich, 2022-11-04, retrieved 2022-11-04

8. *Bühlmann, Matthias (2022-09-28). "stable diffusion based image compression". Medium. Retrieved 2022-11-02*