AI Image Generation

Alan B
PSNA College of Engineering and Technology
Dindigul, TamilNadu, India
aalanbabu147@psnacet.edu.in

Arun Pandi S
PSNA College of Engineering and Technology
Dindigul, TamilNadu, India
arunsaravanan096@psnacet.edu.in

Ekhesh Ram P K
PSNA College of Engineering and Technology
Dindigul, TamilNadu, India
ekheshram7593@psnacet.edu.in

Fariz F
PSNA College of Engineering and Technology
Dindigul, TamilNadu, India
ffariz9361@psnacet.edu.in

***Abstract*—Text to image synthesis refers to the method of generating images from the input text automatically. Deciphering data between picture and text is a major issue in artificial intelligence. Automatic image synthesis is highly beneficial in many ways. Generation of the image is one of the applications of conditional generative models. For generating images, GAN(Generative Adversarial Models) are used. Recent progress has been made using Generative Adversarial Networks (GAN).**

Keywords—GAN,Decipher

INTRODUCTION

Artificial intelligence was founded as an academic discipline in 1956, and in the decades since has experienced several waves of optimism. Since its founding, researchers in the field of artificial intelligence have raised philosophical arguments about the nature of the human mind and the ethical consequences of creating artificial beings endowed with human-like intelligence; these issues have previously been explored by myth and philosophy since antiquity.

AI image generators can produce high-quality art and realistic images much faster than humans. Everyone can create images and artwork without having to know technical drawing skills. A designer or artist will focus more on the concept and imagination of an image than on its technical aspects. This form of art will enable humanity to unlock the unexplored sides of imagination with the help of next level of intelligence created by mankind for the betterment and development of human race towards the future with strong ability to sustain humanity in this vasteness of the observable universe.

1. Justification

1) They can help artists develop ideas for new works they need to do. For example, a concept artist can create many ideas for a video game in less than an hour. Everyone can create images and artwork without having to know technical drawing skills. AI image generators can quickly and easily generate large numbers of images, making them ideal for marketing applications.

2) Many mechanisms for creating AI art have been developed, including procedural "rule-based" generation of images using mathematical patterns, algorithms which simulate brush strokes and other painted effects, and artificial intelligence or deep learning algorithms, such as generative adversarial networks (GANs) and transformers.One of the first significant AI art systems is Arron, developed by Harold beginning in the late 1960s at the University of California at San Diego. AARON is the most notable example of AI art in the era of Gofai programming because of its use of a symbolic rule-based approach to generate technical images. Cohen developed AARON with the goal of being able to code the act of drawing. In its primitive form, AARON created simple black and white drawings. Cohen would later finish the drawings by painting them. Throughout the years, he also began to develop a way for AARON to also paint. Cohen designed AARON to paint using special brushes and dyes that were chosen by the program itself without mediation from Cohen.

3) Generative adversarial networks (GANs) were designed in 2014. This system uses a "generator" to create new images and a "discriminator" to decide which created images are considered successful. More recent models use Vector Quantized Generative Adversarial Network and Contrastive Language–Image Pre-training.

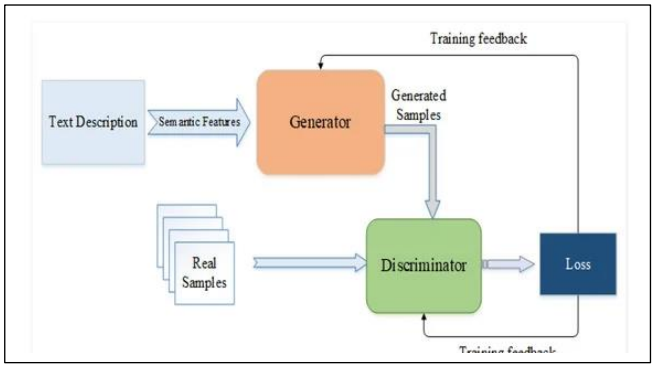
IMPLEMENTATION

Stable Diffusion is a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, out painting, and generating image-to-image translations guided by a text prompt. It was developed by the start-up Stability AI in collaboration with a number of academic researchers and non-profit organizations.

Stable Diffusion is a latent diffusion model, a kind of deep generative neural network. Its code and model weights have been released publicly, and it can run on most consumer hardware equipped with a modest GPU with at least 8 GB VRAM. This marked a departure from previous proprietary text-to-image models such as DALL-E and Mid journey which were accessible only via cloud services.

A table heading (using the “table head” style) appears above a table. This will automatically number the table for you. Any footnotes appear below the table, using the “table footnote” style. Footnotes are indicated by superscript

lowercase letters within the table. An example of a table can be seen in Table I, below.



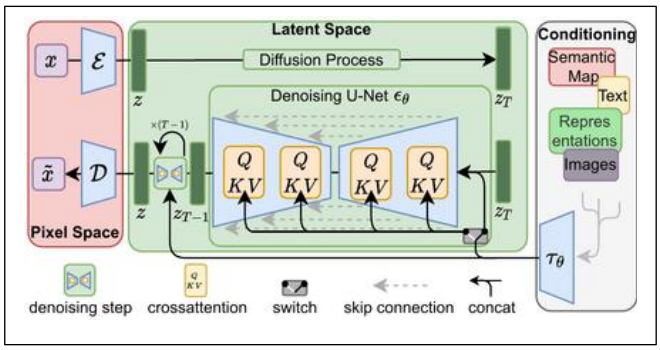
Development

The development of Stable Diffusion was funded and shaped by the start-up company Stability AI. The technical license for the model was released by the Comp Vis group at Ludwig Maximilian University of Munich. Development was led by Patrick Esser and Robin Rombach, who were among the researchers who had earlier invented the latent diffusion model architecture used by Stable Diffusion. Stability AI also credited Eleuther AI and LAION (a German nonprofit which assembled the dataset on which Stable Diffusion was trained) as supporters of the project. In October 2022, Stability AI raised US$101 million in a round led by Lightspeed Venture Partners and Coatue Management.

Architecture

Stable Diffusion uses a kind of diffusion model (DM), called a latent diffusion model (LDM) developed by the Comp Vis group at LMU Munich. Introduced in 2015, diffusion models are trained with the objective of removing successive applications of Gaussian noise on training images which can be thought of as a sequence of denoising autoencoders.

Stable Diffusion consists of 3 parts: the variational autoencoder (VAE), U-Net, and an optional text encoder. The VAE encoder compresses the image from pixel space to a smaller dimensional latent space, capturing a more fundamental semantic meaning of the image.



Gaussian noise is iteratively applied to the compressed latent representation during forward diffusion. The U-Net block, composed of a Res Net backbone, denoises the output from forward diffusion backwards to obtain a latent representation.

Finally, the VAE decoder generates the final image by converting the representation back into pixel space. The denoising step can be flexibly conditioned on a string of text, an image, or another modality.

The encoded conditioning data is exposed to denoising U-Nets via a cross-attention mechanism. For conditioning on text, the fixed, pretrained CLIP VIT-L/14 text encoder is used to transform text prompts to an embedding space. Researchers point to increased computational efficiency for training and generation as an advantage of LDMs.

Training data

Stable Diffusion was trained on pairs of images and captions taken from LAION-5B, a publicly available dataset derived from Common Crawl data scraped from the web, where 5 billion image-text pairs were classified based on language and filtered into separate datasets by resolution, a predicted likelihood of containing a watermark, and predicted "aesthetic" score (e.g. subjective visual quality).

The dataset was created by LAION, a German non-profit which receives funding from Stability AI. The Stable Diffusion model was trained on three subsets of LAION-5B: laion2B-en, laion-high-resolution, and laion-aesthetics.

A third-party analysis of the model's training data identified that out of a smaller subset of 12 million images taken from the original wider dataset used, approximately 47% of the sample size of images came from 100 different domains, with Pinterest taking up 8.5% of the subset, followed by websites such as WordPress, Blogspot, Flickr, DeviantArt and Wikimedia Commons.

Text to image generation

The text to image sampling script within Stable Diffusion, known as "txt2img", consumes a text prompt in addition to assorted option parameters covering sampling types, output image dimensions, and seed values. The script outputs an image file based on the model's interpretation of the prompt. Generated images are tagged with an invisible digital watermark to allow users to identify an image as generated by Stable Diffusion, although this watermark loses its efficacy if the image is resized or rotated.

Each txt2img generation will involve a specific seed value which affects the output image. Users may opt to randomize the seed in order to explore different generated outputs, or use the same seed to obtain the same image output as a previously generated image. Users are also able to adjust the number of inference steps for the sampler; a higher value takes a longer duration of time, however a smaller value may result in visual defects.

Another configurable option, the classifier-free guidance scale value, allows the user to adjust how closely the output image adheres to the prompt. More experimentative use cases may opt for a lower scale value, while use cases aiming for more specific outputs may use a higher value.

Additional text2img features are provided by front-end implementations of Stable Diffusion, which allow users to modify the weight given to specific parts of the text prompt. Emphasis markers allow users to add or reduce emphasis to keywords by enclosing them with brackets.

An alternative method of adjusting weight to parts of the prompt are "negative prompts". Negative prompts are a feature included in some front-end implementations, including Stability AI's own Dream Studio cloud service, and allow the user to specify prompts which the model should avoid during image generation.

The specified prompts may be undesirable image features that would otherwise be present within image outputs due to the positive prompts provided by the user, or due to how the model was originally trained, with mangled human hands being a common example.

Image modification

Stable Diffusion also includes another sampling script, "img2img", which consumes a text prompt, path to an existing image, and strength value between 0.0 and 1.0. The script outputs a new image based on the original image that also features elements provided within the text prompt. The strength value denotes the amount of noise added to the output image. A higher strength value produces more variation within the image but may produce an image that is not semantically consistent with the prompt provided.

The ability of img2img to add noise to the original image makes it potentially useful for data anonymization and data augmentation, in which the visual features of image data are changed and anonymized.

The same process may also be useful for image upscaling, in which the resolution of an image is increased, with more detail potentially being added to the image.

Additionally, Stable Diffusion has been experimented with as a tool for image compression. Compared to JPEG and WebP, the recent methods used for image compression in Stable Diffusion face limitations in preserving small text and faces.

Additional use-cases for image modification via img2img are offered by numerous front-end implementations of the Stable Diffusion model. Inpainting involves selectively modifying a portion of an existing image delineated by a user-provided layer mask, which fills the masked space with newly generated content based on the provided prompt.

A dedicated model specifically fine-tuned for inpainting use-cases was created by Stability AI alongside the release of Stable Diffusion 2.0. Conversely, out painting extends an image beyond its original dimensions, filling the previously empty space with content generated based on the provided prompt.

FUTURE ENCHANECMENT

Additional functionalities are under development and may improve various applications or enable new ones – such as "Textual Inversion" which refers to enabling the use of user-provided concepts (like an object or a style) learned from few images. With textual inversion, novel personalized art can be generated from the associated word(s) (the keywords that have been assigned to the learned, often abstract, concept) and model extensions/fine-tuning (see also: Dream Booth).

Generated images are sometimes used as sketches or low-cost experimentations or illustration of proof-of-concept -stage ideas – additional functionalities or improvements may also relate to post-generation manual editing (polishing or artistic usage) of prompts-based art additional citation(s) needed (such as subsequent tweaking with an image editor). In the case of Stable Diffusion, the main pre-trained model is shared on the Hugging Face Hub.

REFERENCES

[1] https://en.wikipedia.org/wiki/Stable_Diffusion.

[2] https://platform.openai.com/docs/introduction.

[3] https://platform.stability.ai/?_gl=1*14375rf*_ga*MTE2MjIwMDIwNi4xNjcyNjY1ODg1*_ga_W4CMY55YQZ*MTY4MDg2MTQwOC41LjAuMTY4MDg2MTQwOC4wLjAuMA.. .

[4] https://stability.ai/blog/stablediffusion2-1-release7-dec-2022.

[5] Chambon, Pierre; Bluethgen, Christian; Langlotz, Curtis P.; Chaudhari, Akshay (2022-10-09). "Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains"