Chapter 7. Introduction to Diffusion Models for Image Generation

A NOTE FOR EARLY RELEASE READERS

With Early Release ebooks, you get books in their earliest form—the authors' raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the 7th chapter of the final book. Please note that the GitHub repo will be made active later on.

If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the author at ccollins@oreilly.com.

This chapter introduces the most popular diffusion models for AI image generation. You'll learn the benefits and limitations of each of the top models, so that you can be confident in choosing between them based on the task at hand.

Introduced in 2015, *diffusion models* are a class of generative models that have shown spectacular results for generating images from text. The release of DALL-E 2 in 2022 marked a great leap forward in the quality of generated images from diffusion models, with open-source Stable Diffusion, and community favorite Midjourney quickly following to forge a competitive category. With the integration of DALL-E 3 into ChatGPT, the lines will continue to blur between text and image generation. However, advanced users will likely continue to require direct access to the underlying image generation model, to get the best results.

Diffusion models are trained by many steps of adding random noise to an image, and then predicting how to reverse the diffusion process by *denoising* (removing noise). The approach comes from physics, where it has been used for simulating how particles *diffuse* (spread out) through a medium. The predictions are conditioned on the description of the image, so if the resulting image doesn't match, the neural network weights of the model are adjusted to make it better at predicting the image from the description. When trained, the model is able to take random noise and turn it into an image that matches the description provided in the prompt. Figure 7-1 illustrates the denoising process, as demonstrated by Binxu Wang in "Mathematical Foundation of Diffusion Generative Models".

Figure 7-1. Diffusion schematics

These models were trained on large datasets of billions of images scraped from the internet (and accompanying captions), and can therefore replicate most popular art styles or artists. This has been the source of much controversy, as copyright holders seek to enforce their legal claims, while model creators argue in favor of fair use.

A diffusion model is not simply a "complex collage tool" that regurgitates replicas of copyrighted images: it's only a few gigabytes in size and therefore can't possibly contain copies of all its training data. When researchers attempted to reproduce 350,000 images from Stable Diffusion's training data, they only succeeded with 109 of them (Carlini et al, 2023). What the model is doing is more analogous to a human artist looking at every image on the internet and learning the patterns that define every subject and style. These patterns are encoded as a *vector representation* (a

list of numbers) referring to a location in *latent space*: a map of all possible combinations of images that could be generated by the model. The prompt input by the user is first encoded into vectors, then the diffusion model generates an image matching these vectors, before the resulting image is decoded back into pixels for the user. Figure 7-2 illustrates the encoding and decoding process, from Ian Stenbit's "A walk through latent space with Stable Diffusion":

Figure 7-2. Encoding and decoding process

These vectors, also referred to as *embeddings*, act as a location or address for a point in the model's map of every image, and as such images that are similar will be close together in latent space. The latent space is continuous and you can travel between two points (interpolate) and still get valid images along the way. For example, if you interpolate from a picture of a dog to a bowl of fruit, the intermediate images will be coherent-looking images, demonstrating a progressive shift between the two concepts. Figure 7-3 contains a grid, also from Ian Stenbit, showing the intermediate steps between four images: a dog (top left), a bowl of fruit (top right), the Eiffel Tower (bottom left), and a skyscraper (bottom right):

Figure 7-3. A random walk through latent space

Within the domain of diffusion models, prompt engineering can be seen as navigating the latent space, searching for an image that matches your vision, out of all of the possible images available. There are many techniques and best practices for locating the right combination of words to conjure up your desired image, and an active community of AI artists and researchers have worked to build a set of tools to help. Each model and method has its own quirks and behaviors depending on its architecture, training method, and the data on which it was trained. The three main organizations responsible for building the most popular text-to-image diffusion models have all taken radically different approaches in terms of business models and functionality, and as such there is a greater diversity of choice in diffusion models than there is in the OpenAI-dominated large language model (LLM) space.

OpenAI DALL-E

In January 2021 OpenAI released the text-to-image model DALL-E, its name being a play on surrealist artist Salvador Dali and the Pixar animated robot WALL-E. The model was based on a modified version of OpenAI's remarkable GPT-3 text model, which had been released seven months before. DALL-E was a breakthrough in generative AI, demonstrating artistic abilities most people thought were impossible for a computer to possess. Figure 7-4 shows and example of the first version of DALL-E's capabilities:

Figure 7-4. DALL-E capabilities

The DALL-E model was not open-sourced nor released to the public but it inspired multiple researchers and hobbyists to attempt to replicate the research. The most popular of these models being DALL-E Mini, released in July 2021 (renamed Craiyon a year later at the request of OpenAI), and although it gained a cult following on social media, the quality was considerably poorer than the official DALL-E model. OpenAI published a paper announcing DALL-E 2 in April 2022, and the quality was significantly higher, attracting a waitlist of one million people. Figure 7-5 shows an example of the now iconic Astronaut riding a horse image from the paper that captured the public's imagination:

Figure 7-5. DALL-E 2 image quality

Access was limited to waitlist users until September 2022, due to concerns about AI ethics and safety. Generation of images containing people was initially banned, as were a long list of sensitive words. Researchers identified DALL-E adding the words *black* or *female* to some image prompts like `a photo of a doctor` in a hamfisted attempt to address bias inherited from the dataset (images of doctors on the internet are disproportionally of white males).

The team added Inpainting and Outpainting to the user interface in August 2022, which was a further leap forward, garnering attention in the press and on social media. These features allowed users to generate only selected parts of an image, or to *zoom out* by generating around the border of an existing image. However, users have little control over the parameters of the model, and could not fine-tune it on their own data. The model would generate garbled text on some images, and struggled

with realistic depictions of people, generating disfigured or deformed hands, feet, and eyes, demonstrated in Figure 7-6:

Figure 7-6. Deformed hands and eyes

Google's Imagen demonstrated impressive results and was introduced in a paper in May 2022 (Ho et al, 2022), but the model was not made available to the general public, citing AI ethics and safety concerns. Competitors like Midjourney (July 2022) moved quickly and capitalized on huge demand from people who had seen impressive demos of DALL-E on social media, but were stuck on the waitlist. The open-source release of Stable Diffusion (August 2022) broke what had seemed to be an unassailable lead for OpenAI just a few months before. Although the rollout of the more advanced DALL-E 3 model as a feature of ChatGPT has helped OpenAI re-gain lost ground, and Google have gotten into the game with Gemini 1.5, there remains everything to play for.

Midjourney

In July 2022, just three months after the release of DALL-E 2, Midjourney put its v3 model in open beta. This was a uniquely good time to launch an image generation model, because the demonstrations of what DALL-E 2 could do from early users looked like magic, and yet access was initially limited. Eager early-adopters flocked to Midjourney, and its notable fantasy aesthetic gained a cult following among the gaming and digital art crowds, showcased in the now famous image, which won first prize in a digital art competition, in Figure 7-7:

Figure 7-7. Théâtre d'Opéra Spatial

Midjourney was one of the first viable image models that had a business model and commercial license, making it suitable for more than just experimentation. The subscription model was favored by many artists accustomed to paying monthly for other software like Adobe Photoshop. It also helped the creative process to not be charged per image generated, particularly in the early days when you'd have to try multiple images before you found one that was high enough quality. If you were a paying customer of Midjourney you owned the rights to any image generated, unlike DALL-E where OpenAI was retaining the copyright.

Unique to Midjourney is its heavy community focus. In order to use the tool you must sign into a Discord server (Figure 7-8) and submit your prompt in an open channel or direct message. Given that all image generations are shared in open channels by default, and private mode is only available on the most expensive plan, the vast majority of images created through Midjourney are available for others to learn from. This led to rapid copying and iteration between users, making it easy for novices to quickly learn from others. As early as July 2022 the Discord community was nearing one million people (shown in Figure 7-8, and a year later there were over 13 million members.

Figure 7-8. Midjourney's Discord server July 2022

When you find an image you like, you can click a button to *upscale* the image (make it higher resolution) for use. Many have speculated that this procedure acts as training data for reinforcement learning, similar to Reinforcement Learning from Human Feedback (RLHF), the method touted as the key to success of ChatGPT. In addition, the team regularly asks for ratings of images generated by newer models in order to improve the performance. Midjourney released v4 of its model in November 2022, followed by v5 in March 2023 and v6 in December 2023. The quality is significantly improved, hands and eyes issues identified in Figure 7-6 have largely gone away, and the model has a larger stylistic range, demonstrated in Figure 7-9:

Input:

```
a group of best friends women eating salads and laughing
while high fiving in a coffee shop, cinematic lighting
```

The output is shown in Figure 7-9.

Figure 7-9. Women eating salads and laughing

Remarkably the Midjourney team has remained small, with just 11 employees as of March 2023. The founder of Midjourney, David Holz, formerly of hardware startup Leap Motion, confirmed in an interview that the company was already profitable as of August 2022. What is even more remarkable, is that without the billions of dollars of funding that OpenAI enjoys, the team has built significant functionality over what's available in DALL-E, including negative prompting (removing concepts from an image), weighted terms (increasing the prevalance of other concepts), and their *describe* feature (reverse-engineering the prompt from an uploaded image). However, there is no API available and the only way to access the model is through Discord, which has likely acted as a drag on mainstream adoption.

Stable Diffusion

While DALL-E 2's waitlist continued to build, researchers from the CompVis Group at LMU Munich and applied research company Runway ML received a donation of computing power from Stability AI to train Stable Diffusion. The model shocked the generative AI world when it was released open source in August 2022, because the results were comparable to DALL-E 2 and Midjourney, but it could be run for free on your own computer (assuming you had a modest GPU with 8GB VRAM). Stable Diffusion had one of the fastest climbs in GitHub stars of any software, rising to 22,000 stars in less than 2 months (Figure 7-10):

Figure 7-10. GitHub developer adoption of Stable Diffusion

The move to open-source the model was controversial, and raised concerns about AI ethics and safety. Indeed, many of the initial use cases were to generate AI porn, as evidenced by the Not Safe For Work (NSFW) models shared on platforms like Civitai. However, the ability for hobbyists and tinkerers to modify and extend the model, as well as fine-tune it on their own data, led to rapid evolution and improvement of the model's

functionality. The decision to surface all of the model's parameters to users, such as Classifier Free Guidance (how closely to follow a prompt), Denoising (how much noise to add to the base image for the model to remove during inference), and Seed (the random noise to start denoising from), has led to more creativity and innovative artwork. The accessibility and reliability of open-source has also enticed several small businesses to build on top of Stable Diffusion, such as Pieter Level's PhotoAI and InteriorAI (together raking in over $100,000 in monthly revenue), and Danny Postma's HeadShot Pro. As well as matching DALL-E's inpainting and outpainting functionality, open-source contributions have also kept pace with Midjourney's features, such as negative prompts, weighted terms, and the ability to reverse-engineer prompts from images. In addition, advanced functionality like ControlNet (matching the posture or composition of an image) and Segment Anything (clicking on an element to generate a mask for inpainting), have been quickly added as extensions for use with Stable Diffusion (both released in April 2023), most commonly accessed via AUTOMATIC1111's Web UI (Figure 7-11):

Figure 7-11. AUTOMATIC1111's Web User Interface for Stable Diffusion

Version 1.5 of Stable Diffusion was released in October 2022, and is still in use today. Therefore it will form the basis for the ControlNet examples in Chapter 10, the advanced section for image generation in this book. The weights for Stable Diffusion were released on HuggingFace, introducing a generation of AI engineers to the open-source AI model hub. Version 2.0 of Stable Diffusion came out a month later in November 2022, trained on a more aesthetic subset of the original LAION-5B dataset (a large-scale dataset of image and text pairs for research purposes), with NSFW (Not Safe For Work) images filtered out. Power users of Stable Diffusion complained of censorship as well as a degradation in model performance, speculating that NSFW images in the training set were necessary to generate realistic human anatomy.

Stability AI raised over $100m and has continued to develop newer models, including DeepFloyd, a model better able to generate real text on images (an issue that plagues other models) and the current favorite Stable Diffusion XL 1.0 (abbreviated to SDXL). This model has overcome the misgivings of the community over censorship in version 2.0, not least due to the impressive results of this more powerful model, which has 6.6 billion parameters, compared with 0.98 billion for the v1.5 model.

Google Gemini

Google long threatened to be a competitor in the space with their Imagen model (not released publicly), and indeed ex-Googlers have since founded a promising new image model Ideogram, released in August 2023. They finally entered the image generation game with Gemini in December 2023, though quickly faced criticism over a clumsy attempt to promote diversity. It remains to be seen whether Google's internal politics will prevent them from capitalizing on their significant resources.

Text to Video

Much of the attention in the image space is also likely to shift towards *text-to-video*, *image-to-video* and even *video-to-video*, as the Stable

Diffusion community extends the capabilities of the model to generate consistent images frame-by-frame, including promising open-source projects such as AnimateDiff. In addition, one of the co-creators of Stable Diffusion, RunwayML has become the leading pioneer in text-to-video, and is starting to get usable results with their Gen-2 model. Stable Video Diffusion was released in November 2023, capable of turning text into short video clips or animating existing images, and Stable Diffusion Turbo can generate images in near real time. The release of Sora in February 2024 shows that OpenAI aren't sleeping on this space either. Although we don't cover text to video prompting techniques explicitly, everything you learn about prompting for image generation applies directly to video.

Model Comparison

As demand for AI image generation increases and competition heats up, new entrants will emerge and the major players will diversify. In our own workflows we already find ourselves using different models for different reasons. DALL-E 3 is great at composition, and the integration with ChatGPT is convenient. Midjourney still has the best aesthetics, both for fantasy and photorealism. Stable Diffusion being open-source makes it the most flexible and extendable model, and is what most AI businesses build their products on top of. Each model has evolved towards a distinct style and set of capabilities, as can be discerned when comparing the same prompt across multiple models, as in Figure 7-12:

Input:

```
a corgi on top of the brandenburg gate
```

The output is shown in Figure 7-12.

Figure 7-12. A Corgi on top of the Brandenburg gate

Summary

In this chapter, you were introduced to diffusion models for AI image generation. These models, such as DALL-E, Stable Diffusion, and Midjourney, use random noise and denoising techniques to generate images based on text descriptions. They have been trained on large datasets and can replicate various art styles. However, there is controversy surrounding copyright issues. You learned how prompt engineering principles apply to image generation when navigating the latent space to find the desired image.

In this chapter you explored the different approaches taken by organizations like OpenAI, Stability AI, and Midjourney in developing text-to-image models. OpenAI's DALL-E gained popularity for its artistic abilities, but access was limited and the quality of replicated models was poorer. Midjourney, on the other hand, capitalized on the demand for DALL-E alternatives and gained a cult following with its v3 model. It had a subscription-based pricing model and a strong community focus. Stable Diffusion, on the other hand, gained attention for its comparable results to DALL-E and Midjourney, but with the advantage of being open source and free to run on personal computers. By reading this chapter, you have also gained insights into the history of AI image generation, and the advancements made by organizations like OpenAI, Midjourney, and Stable Diffusion.

In the next chapter, you will embark on a journey into the world of Image Generation using AI. This chapter will equip you with the necessary knowledge and techniques to create visually stunning and unique images. From format modifiers to art style replication, you will discover the power of prompt engineering in creating captivating and original visual content. Get ready to unleash your creativity and take your image generation skills to new heights.