Chris Hales

Text-to-Image Synthesis of the River Daugava: An Artistic Investigation

ADAMarts
Volume 4, 2023
Audiovisual Media Arts
DOI: 10.32025/adamarts23008
Received: 14.08.2023
Accepted: 01.09.2023

Abstract

The field of AI-generated art has recently burgeoned due to developments in a variety of image-generating neural network techniques. Typing text prompts to produce imagery, a process known as text-to-image, underwent significant performance improvements in late 2022, leading to an uptake in interest and popularity. This paper presents artistic research based around experimentation with image generation representing the River Daugava. Image synthesis outputs are compared when utilising a generative adversarial network (GAN) and the Stable Diffusion text-to-image model, and the importance of the prompt text to the Stable Diffusion results is explored. Information is also presented in terms of the technical aspects of image synthesis and text-to-image, and the ethical and artistic considerations that arise are outlined.

Keywords

artificial intelligence, machine learning, neural network, GAN, text-to-image, Stable Diffusion, prompt engineering

...

Introduction

Developments are occurring continually in the use of artificial intelligence (AI) across a range of societal applications, including those for creative purposes, which are examined in this paper. Neural networks can be trained and deployed using highly complex and processor-intensive programming techniques in a process known as machine learning (ML). Innovations in this field generally occur through the publication of research papers (by commercial entities as well as research scientists) which are immediately examined by both entrepreneurs and the creative community, with websites and apps developed almost overnight to implement the new research. Certain developments may require leveraging a particular dataset or a pretrained model which may, or may not, be made publicly available as open source. A good example of this is the StyleGAN neural network and its various improvements (made available from 2018 onwards) researched and trained by NVIDIA and, more pertinent to this article, the CLIP model released publicly by OpenAI in 2021. Before CLIP, which interprets text guidance, a rudimentary text-to-image model called AttnGAN was described in 2021 as creating images that are "rarely if ever representational and resemble a fusion of abstraction and Post-Impressionism" (Hales, 2021, p. 80). Performance improvements since then, as a result of massive increases in amounts of data and the computational power thrown at them, have been nothing less than astonishing. This article aims, by means of a qualitative approach that does not stray into computer science, to examine outputs from image-based ML models in two forms: as generative adversarial networks (GANs), which can be self-trained using personally collected visual imagery, and recently developed text-to-image models such as *Stable Diffusion* (hereinafter called SD), which have been trained on massive amounts of data taken from online sources. By the process of generating Latvia-specific image outputs, a

comparison can be made between a custom-made model (created by the author of this article) and one that has been trained on a vast worldwide cultural database, and the effects of varying a text prompt to guide the latter model can be explored. At the time of writing, the amount of published academic literature directly relevant to this discussion is surprisingly scarce, the bulk of publications being online journalistic articles or scientific papers dealing with computational techniques. The publication time lag struggles to keep up with the almost daily developments in the techniques and quality of AI-generated imagery, and publications relevant to this article are either yet to appear or are already outdated. To overcome this, Manovich and Arielli have been publishing a book incrementally online since 2021 to document what Manovich terms an 'ongoing revolution' (Manovich, 2023). By contrast, Arthur Miller's *The Artist in the Machine*, an informative text back in 2019, already seems like a history book: most of the case studies and techniques presented have already been superseded or have lost favour as researchers introduce newer types of implementation with higher image resolution and quality. Face swapping, for example, was a mobile phone app craze around 2016, and shortly afterwards neural style transfer became popular – this being a process in which the higher stylistic layers of a pretrained model (trained for example on paintings by Monet or Van Gogh) are applied to a user-supplied content image. The popularisation of the weird and psychedelic 'DeepDream' imagery also dates to around this time. Although techniques such as these might now be considered little more than gimmicks, by 2019 image synthesis using generative adversarial networks (GANs) provided a method to originate imagery based on a user-supplied dataset that could be trained using ready-made code. At the time of writing, however, a recently developed alternative approach to image synthesis known as diffusion has taken centre stage; datasets used for training have increased dramatically in size and scope; and text guidance techniques have developed as a result of CLIP. This has led to a popularisation of the use of text guidance to generate image (or text) output, with the term 'text-to-image' becoming a predominant theme in the AI-generated art community. A proliferation of websites and apps emerged in the last quarter of 2022 and early 2023 to implement these text-to-image developments: *Craiyon, Dream by Wombo, Midjourney, NightCafé* and Stability AI's *DreamStudio* are currently some of the most popular. The terms 'AI-generated art' and 'generative art' are now widely used to categorise the type of imagery, very often fantasy-inspired and photorealistic, which is generated by use of a text-to-image model. It must be noted, however, that practices in which aspects of the creative process are relinquished to an external mechanism are not new (Kelomees, 2022) and generative art is an established field that covers a great many techniques and styles, including abstraction. Given the paucity of relevant academic literature, a methodology of action research via experimentation is adopted. The author of this paper firstly gained the expertise to create several unique GANs, including one trained on imagery of the Daugava River, chosen because it is a specific Latvian subject that might potentially be absent from the huge dataset used to create *Stable Diffusion* – thereby revealing a difference in capability between SD and GAN methods. Experimentation compared sample outputs from both approaches in order to make a subjective analysis of how both models generate the same target output (Daugava). Subsequently, in order to reveal variation produced by the text prompt, the study investigated the additional capability of SD to extemporise on the basic visual theme. Finally, the text prompts were further elaborated to reveal the potential for both faking reality and encouraging creative imagination when using text-to-image generators. This approach is limited by very small amounts of data (in the form of the generated images) and the fact that results are subjectively commented upon only by the author.

The aim nevertheless is to better understand the processes behind GANs and text-to-image generators and the affordances they offer when used for creative purposes.

How Text-to-Image Works

Image-generating neural networks are trained over a substantial number of iterations (sometimes called epochs) from a large dataset of images which may be original photographs or artworks, or 'scraped' from the Internet. As the training process progresses, information is gradually built into a neural network that begins to understand lines and shapes and eventually learns how to create new and original images that resemble, but are different from, images that were in the original dataset. If the neural network is not given any description of each item in the dataset, then the process is described as unsupervised. In supervised learning, categories (classes) are assigned to each image (e.g., this image is a dog, this image is a cat), meaning that a model could be trained to perform as a 'classifier' which could try to identify previously unseen images; alternatively, it could generate new exemplars of each particular class on which it was trained. Another approach is to train pairs of images so as to produce a model that could perform an image-to-image transformation; for example, a high-resolution image could be paired with a low-resolution image of the same, such that the model learns to upscale an image if only the low-resolution version is supplied. Most importantly for text-to-image generation, image+caption pairs comprise the dataset: each image is paired with a text description of that image. Since it would be tedious (though not impossible) for anyone to try to create a useable dataset this way, the images are invariably scraped from websites across the Internet along with their accompanying captions. In this way, any image posted online with a caption, for example on a personal blog, stock image website, or a news or sports website, has the potential to end up in such a dataset and be used in the training process, thereby influencing the outputs produced. The CLIP text guidance model which brought about the current rapid developments in text-to-image was trained by the company Open-AI on such a massive dataset and released in early 2021.
What CLIP (and its variations) does in the simplest possible terms is to iteratively compare an image and a caption and report the accuracy of the match. CLIP deals with the text guidance and does not itself create the imagery: it is tightly coupled with an image generator model (which could be a GAN but is more commonly a diffusion model trained to refine a pattern of random noise into meaningful images) to make up the entire package, and these combinations often have catchy titles such as *Aphantasia, Hypertron, Disco Diffusion* and OpenAI's *DALL-E*. The aim of the process is to find the best possible match between a generated image and the user-supplied text descriptions: images are repeatedly generated in an iterative series of steps, each of which gets a score from the text guidance model with the aim of continually improving the match with the text prompt until a certain number of steps (for example, fifty) have been completed. Because this process will result in a different image each time, a batch of varying output images can be generated from which the user can cherry-pick those images that are considered most successful. At the time of writing, *Stable Diffusion* (SD) developed by Stability AI is an overwhelmingly popular community-driven solution that is considered to give realistically rendered outputs that excel in matching image to prompt – the text-to-image experiments later in this article will be based on the use of SD. Although its first version (August 2022) used the original CLIP model for the text guidance, the second version (November 2022) has its own text guidance model (OpenCLIP) trained using a 120-million-image subset of the LAION-5B dataset of 5.85 billion image-caption pairs (filtered for inappropriate content and aesthetic value). SD is able to combine quite disparate concepts in original ways, such that the

results generated from a prompt like, for example, "a car in the shape of a dog" go far beyond mere collage and are often intriguing and ingenious in their unexpected inventiveness. It is for this reason that SD and comparable models are considered by some to show an almost magical advancement in image-based generative AI – bearing in mind of course that the 'magic' originates in the massive corpus of human ingenuity contained in the datasets upon which the models have been trained. Not all text-to-image models build their final output exclusively on text guidance: visual guidance can also be used as a complementary tool to craft a desired image. Alternative solutions exist for this, the most direct approach being one which affords the user the possibility to upload a starting image from which the text-to-image iteration is commenced. Artbreeder's *Collager* system has a different methodology that allows the user to position some basic shapes that suggest where particular prompt elements should be drawn, and Meta AI's *Make-A-Scene* offers a basic sketching interface with a user's drawings providing visual guidance to supplement the work of the text prompts.

DaugavaGAN and Daugava Stable Diffusion

Inherent in the GAN-type architecture referred to earlier is the narrow specialism as to what the model can create. If trained on a few thousand photographs of cars, it would only produce attempts at new car images, and so on. The author of this article has experience training several GAN models on specific image datasets collated by himself, primarily through scraping images from Instagram: this has resulted in a neural network that can generate images of Baltic mitten designs; a network that generates images of the Hill of Crosses in Šiauliai, Lithuania; and one that creates visual imagery of the River Daugava in Latvia. In each case, these were probably the first image-based neural networks ever created on that particular subject. There are a great number of these specialised GAN models either held privately (as my own) or made publicly available for others to use. This narrow functionality reveals a situation which is far away from the goal of artificial general intelligence (AGI), in which the neural network becomes versatile and multifunctional enough to begin to rival human intelligence itself. *Stable Diffusion*, however, does make a slight but significant move in that direction because the enormous dataset on which it was trained broadens the range of imagery that can be produced.
Below are reproduced a variety of 512px square images generated from the author's neural network model of the River Daugava (named *DaugavaGAN* by the author) which have been generated randomly. The original dataset consisted of 1,000 images of the Daugava at different times of day and at different locations, and the training took 340 hours of GPU compute time. Additional training steps and a larger dataset would have been desirable to improve the results; nevertheless, features such as the railway bridge and the support pillar of the Vanšu tilts are quite recognisable, albeit in a somewhat freeform representation. There was no aim to make the output an exact reconstruction but rather to create artistic imagery that could be formed into an experimental film by interpolation through the latent space of the model in the process of 'spacewalking' referred to above. The training was unsupervised, and the network creates imagery without knowledge that this is the River Daugava in Riga (and upstream). Stan Brakhage's description of experimental film as "an adventure of perception" (Brakhage, 1963, "Metaphors on Vision", para. 1) is particularly appropriate to the exploration of the latent space of a GAN, which contains almost unlimited visual variations on the overall theme without any preconception of what the theme represents to a human. The *DaugavaGAN* model does just the one job but with enough fidelity for anyone familiar with the Daugava to sense an impression of familiarity in many of the images the model produces.
Let us now study the *Stable Diffusion* (SD) model

Figure 1. A variety of images generated by the author's DaugavaGAN image-based neural network.

as it stands in early 2023, which is, essentially, a single neural network with knowledge of a great many categories of imagery rather than just the one. Surely the SD text-to-image model could not generate images with the same quality as *DaugavaGAN*? Inputting the text prompts "River Daugava" or "River Daugava in Riga, Latvia" to SD gives 512px square images such as those illustrated below in Figure 2 (these have been selected from amongst multiple possibilities generated by SD). The leftmost two images in Figure 2 are generic river views that might pass for the Daugava upstream from Riga, but the three images on the right of Figure 2 seem to be fantasy constructions with generic spires and red roofs with the topology all wrong. Sample images from the same prompt can be generated endlessly by SD with outputs undoubtedly hit or miss, meaning that with sheer persistence it may be possible to produce an image that is more recognisable as actually being the Daugava in Riga; nevertheless, a stranger to Riga using SD with these prompts would most likely get a skewed view of reality. As the technology currently stands, it is clear that *DaugavaGAN* outshines SD in its photorealistic representations on this particular theme, although significant time and effort was expended in creating *DaugavaGAN* specifically for this task and it only took a few words of typed text to generate from the SD model. These results beg the question of whether the SD model has any actual knowledge of the River Daugava at all, or whether it is inventing generic scenes. Fortunately, this question can be easily solved because the LAION-5B image+caption dataset on which SD was trained is open-sourced and can be inspected, a website offering this being *haveibeentrained.com*. Entering "River Daugava" into the search box of this website reveals that about 250 photographs are present, although several of these certainly do not represent the Daugava. There are plenty of photographs in the LAION-5B dataset for "Riga, Latvia" and although searching with "RISEBA University" yields a few images, none of these are actually of RISEBA (one possible explanation is that a scraped image of Paris, for example, might have been captioned "students from RISEBA University on a field trip to Paris"). There is evidently excessive leeway in the search engine of *haveibeentrained.com*, but still, we can reasonably conclude that the SD model does indeed have some visual knowledge of the River Daugava and knows about Riga (but does not know anything about RISEBA). The images being generated by SD, therefore, are clearly not mere collages cut-and-pasted from the samples in the dataset but are original compositions freely extemporised from the knowledge embedded in the SD neural network. A brief study of online galleries of SD outputs quickly reveals that it is being used overwhelmingly to create fantastical scenes and imaginative combinations in a hyperrealistic manner rather than to reconstruct reality itself, meaning that SD is undoubtedly best used when allowed to freely undertake its own

Figure 2. Images generated by Stable Diffusion with prompt "River Daugava" (first three images) and "River Daugava in Riga, Latvia" (the two images on the right).

'adventure of perception'.
Where the SD method comes into its own, therefore, is when the text guidance is used to go beyond a basic representation of the core prompt. Some gradual steps in this direction are shown below in Figure 3.
The prompt for the image on the left was "River Daugava, Riga, at night", and the adjacent image prompt was "Railway Bridge over the River Daugava". The representation of nighttime is excellent, although the Riga skyline is wrong, and the railway bridge is also incorrectly represented. The third and fourth images in Figure 3 relate to prompts "water-colour painting of the River Daugava in winter" and "pencil sketch of River Daugava", revealing that if the core prompt can be correctly interpreted by SD, then weather conditions, times of day and stylistic rendering can be carried out with ease. This could be considered a new variety of neural style transfer in which the user need only supply the text guidance rather than uploading a content image. The image on the right of Figure 3 is somewhat worrying in the manner in which a fake narrative has been represented: the prompt was "River Daugava in Riga, floods". This particular image is clearly not Riga, but with persistence a more convincing result could eventually be generated from a similar prompt. Moving towards the fantastical, experimenting with the prompt "blurry black-and-white photograph of UFO hovering over Riga city centre" (not illustrated here) can also give believable imagery – untainted by Photoshop retouching – which might easily be passed off as (fake) news. The ease with which believable imagery can be generated combined with the realism of SD has raised debate about the moral and ethical questions around the use of text-to-image. An additional aspect causing concern is that the LAION-5B dataset contains many images scraped from the online art galleries of active bona fide artists, many of whom are, no doubt, struggling to make a living from their art: SD can freely generate believable imagery in the style of these artists that would be politely called pastiche but more colloquially would be criticised as a rip-off. Technologist Andy Baio addresses this by stating that text-to-image "opens profound questions about the ethics of laundering human creativity" (Baio, 2022).

The Art of Prompt Engineering

As would be evident from the above discussion, the exact phrasing of a prompt is critical to generating a desired image. As well as the core (or raw) prompt, a style description such as "pencil sketch" or "a photograph of" is often given; a format can be suggested such as "a landscape painting of"; and specific artists may be invoked by prompts such as "Rembrandt oil painting of". Prompt weighting, both positive and negative, can be specified on chosen text terms; for example, "trees:-1.0" would result in compositions being created with much less chance of trees appearing in them. In this way, text prompts might easily stretch to fifty

Figure 3. Various images generated by Stable Diffusion based on a core prompt of "River Daugava".

or sixty words and might include negative and positive weighting. The process of crafting these complex prompts has become known as 'prompt engineering', a totally new field and a profession which has emerged almost overnight. The fact that significant experimentation is required to discover and refine a particular prompt should not come as a surprise given that the text guidance is the predominant interface by which the generated imagery can be customised and the dataset is based on a vast variety of image+caption pairs. Painstaking trial and error has led many prompt engineers to discover how to generate impressive images: in this way, it was discovered that adding "rendered with Unreal Engine" to a prompt was invaluable to generate hyperrealistic outputs. Although many such prompts are shared with the community, others are kept as a closely guarded secret. The prompt engineer also now has an opportunity to monetise their skill via websites such as *promptbase.com*, which offers a 'prompt marketplace' for buying and selling prompts for a variety of popular text-to-image generators including SD.

Skill with prompt engineering begins with an understanding that the dataset comprises scraped image+caption pairs of very diverse origin and hence prompting with a word such as "beautiful" is highly ambiguous: in the dataset, it might have referred to a variety of images such as the human face, a landscape, flowers, or a painting. Hence, the model is going to have difficulty in interpreting the exact usage that the prompter had in mind – an alternative approach might be to add a negative weighting to "ugly" instead. Likewise, there are difficulties with subtleties of language and idioms: "a fork in the road" might easily end up generating images of a culinary item lying on asphalt; "Pikachu on a skateboard" might result in an image of a skateboard with a Pikachu graphic imprinted on it. Help is at hand with deriving the most accurate text prompt to generate a particular desired outcome: prompt matrices have been created showing grids of generated images with the same core prompt combined with different descriptive prompts, and methods have been developed to interrogate the CLIP model in reverse by submitting an image and discovering how CLIP describes it.

It may already be said that in the field of text-to-image art, the term 'artist' has been superseded by the 'prompt engineer'. In the summer of 2022, Jason Allen used the *Midjourney* text-to-image software to win the digital art section of the Colorado State Fair art competition with an image entitled *Théâtre d'Opéra Spatial*. This resulted in significant antipathy from others in the artistic community who claimed that Allen was not an artist and little more than a cheat. Allen was quoted in an article by Roose for the New York Times (Roose, 2022) as saying that "art is dead, dude. It's over. A.I. won. Humans lost". The argument here is that text-to-image is simply a new creative tool that starts with a human idea, and hours of skill and trial and error are required for a prompt engineer to generate images that represent

that idea in eye-catching ways. Although this ar-gument implies that time and effort expended on perfecting a prompt equates to the quality of artistic endeavour, it can reasonably be claimed that imagination and ingenuity is still required to generate striking imagery from a prompt just as it would be from a pencil or a paintbrush.

Conclusions

The experiments with the River Daugava detailed above reveal the potential inaccuracy of the *Stable Diffusion* model to represent a very specific real-life subject when compared with a purpose-trained GAN. Outputs from the model can be misleading because, although they appear realistic, many elements of the visual composition may not match the actual situation. At the same time, SD can decorate and embellish a basic theme to produce im-pressive results where fidelity to a specific subject is not required and it is demonstrably not just a cut-and-paste collage machine. Where text-to-image comes into its own is probably where the reproduction of known reality is not the goal: the latest text-to-image models encourage their prompters to wild flights of imaginative fancy and can transform dreams into visual realities. It becomes clear upon inspecting online galleries of imagery from text-to-image communities (such as that for *Midjourney*) that those with an interest in fantasy and sci-fi are highly productive – the use of text guidance presumably giving many non-artists the opportunity to visualise their thoughts and ideas for the first time. The ease with which a few words of text can transform within a few seconds into a totally unexpected visual representation is wonderfully empowering but can also bring accusations of trivialising the creative process.

Such is the pace of current developments in text-to-image that this article is out of date as soon as it is written. It is already possible to fine-tune the Stable Diffusion model with the addition of only a handful of images (a variety of photographs of a face, for example) and to create new prompt keywords – meaning, for example, that a user's face could be generated into the SD output. Extending text-to-image to create video sequences is already possible, although an issue with the diffusion method is that it is a frame-by-frame still-image process that lacks temporal coherence. Undoubtedly these specific shortcomings will be solved, and text-to-video will develop in its own right utilising massive scraped datasets of online videos – and text guidance of audio will inevitably follow the same development path. In other fields, there is some research into text-to-3D-shape generation (Sanghi et al., 2022), although datasets are smaller since 3D model files are less frequent online. Runway ML are currently promoting a text-guided video editor software which might be termed text-to-video-editing, and *GitHub Copilot* is already up and running using natural language guidance to assist programming in the Python language. AI-based text generation has not been discussed here but made a name for itself in late 2022 with the release of OpenAI's conversational text model entitled *ChatGPT*.

All output from text-to-image represents an Internet-centric view of the world since all image+caption pairs in the dataset have been sourced online. This opens up the generated imagery to any inherent online bias; for example, if a prompt includes the word "school-teacher", it is most often a female teacher who is portrayed. Ethical concerns are being raised about specific aspects of datasets, for example the inclusion of artwork by currently active artists whose style can be pastiched and potentially sold. Copyrighted images such as Disney characters or celebrity faces are being removed from some training datasets due to fears that future litigation will at some stage be instigated to decide the legality of us-ing particular visual imagery in the training of text-to-image neural networks.

This article has focused on creative practice, the production of visual imagery as artxwork. It is important to point out that the creative and editorial industries are already established users of

text-to-text and text-to-image, which have proved an invaluable aid to rapidly visualise new ideas to present to clients. In a matter of seconds, a striking original illustration can be generated for an article rather than using clip art or the services of an illustrator, and money can be saved by no longer employing human photographers, models and makeup artists. This raises the familiar Luddite fear that illustrators, artists and photographers will lose their jobs. Historically there has always been resistance to new technologies like photography that subsequently earn the right to be considered art forms in their own right. Despite the current accusations of cheating, triviality of effort, and appropriation, will the time come when prompt engineering will be added to the panoply of the visual arts to sit alongside painting, drawing, sculpture and photography? Are prompt engineers the artists of the future?

...

References

Baio, A., 2022. Opening the Pandora's Box of AI Art. Available at: https://waxy.org/2022/08/opening-the-pandoras-box-of-ai-art/ (Accessed: 28 January 2023).

Brakhage, S., 1963. Metaphors on Vision. Film Culture special edition issue number 30, fall 1963.

Hales, C., 2021. Artificial Intelligence: The Latent Revolution in Filmmaking. ADAM arts, Vol. 2. RISEBA University, Riga. pp. 72-87.

Kelomees, R., 2022. Concept Transference in Art and AI. Kelomees, R., Guljajeva, V., Laas, O. (eds.), The Meaning of Creativity in the Age of AI. Estonian Art Academy, Tal-linn. pp. 106-122.

Manovich, L., 2023. AI image and Generative Media: Notes on Ongoing Revolution. Ma-novich, L., Arielli, E. (eds.), Artificial Aesthetics, published one chapter at a time on <ma-novich.net>.

Miller, A. I., 2019. The Artist in the Machine. The World of AI-Powered Creativity. MIT Press. Cambridge, Massachusetts.

Roose, K., 2022. An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. New York Times. Available at: https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html (Accessed: 28 January 2023).

Sanghi, A. et al., 2022. CLIP-Forge: Towards Zero-Shot Text-to-Shape Generation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orle-ans, LA, USA. pp. 18582-18592.

...