The Cultivated Practices of Text-to-Image Generation

Jonas Oppenlaender

University of Jyväskylä

jonas.x1.oppenlander@jyu.fi

Abstract

Humankind is entering a novel creative era in which anybody can synthesize digital information using generative artificial intelligence (AI). Text-to-image generation, in particular, has become vastly popular and millions of practitioners produce AI-generated images and AI art online. This chapter first gives an overview of the key developments that enabled a healthy co-creative online ecosystem around text-to-image generation to rapidly emerge, followed by a high-level description of key elements in this ecosystem. A particular focus is placed on prompt engineering, a creative practice that has been embraced by the AI art community. It is then argued that the emerging co-creative ecosystem constitutes an intelligent system on its own — a system that both supports human creativity, but also potentially entraps future generations and limits future development efforts in AI. The chapter discusses the potential risks and dangers of cultivating this co-creative ecosystem, such as the bias inherent in today's training data, potential quality degradation in future image generation systems due to synthetic data becoming common place, and the potential long-term effects of text-to-image generation on people's imagination, ambitions, and development.

Keywords: generative AI, text-to-image generation, prompt engineering, AI art, creativity, human-AI co-creation

Number of words: 5,380

The Cultivated Practices of Text-to-Image Generation

Introduction

Generative artificial intelligence (AI) has taken the world by storm. Using deep generative models, anybody can conjure up digital information from short descriptive text prompts. Text-to-image synthesis, in particular, has become a popular means for generating digital images (Crowson et al., 2022; Rombach, Blattmann, & Ommer, 2022). Millions of people use generative systems and text-to-image services available online, such as Midjourney[1], Stable Diffusion (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2021), and DALL-E 2 (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022), both for professional and recreational uses. With this powerful generative technology at our fingertips, humankind is ushering into a new era — an era in which visual imagery no longer necessarily reflects the effort put into creating the imagery (Oppenlaender, 2022a).

This chapter first gives an overview of the key technical developments that enabled a co-creative ecosystem around text-to-image generation to rapidly emerge and expand in 2021 and 2022. This is followed by a high-level description of key elements in the ecosystem and its practices. A focus is placed on prompt engineering, a method and creative practice that has proven useful in a broad set of application areas, but has been particularly embraced by the community of text-to-image generation practitioners. It is then argued that the creative online ecosystem constitutes an intelligent system on its own — a system that both enables but also potentially limits the creative potential of future generations of humans and machine learning systems. We discuss some potential risks and dangers of cultivating this co-creative ecosystem, such as the threat of bias due to Western ways of seeing encoded in training data, quality degradation due to synthetic data being used for training future generative systems, and the potential long-term effects of text-to-image generation on people's creativity, imagination, and development.

[1] https://www.midjourney.com

Background on text-to-image generation

The history of computer-generated art and “generative art” (Boden & Edmonds, 2009; Galanter, 2016) goes back to first experiments with AI (Cohen, 1979). Looking back, the first attempts to synthesize images from text were humbling, but already showed great promise. The synthetic images presented by Mansimov, Parisotto, Ba, and Salakhutdinov (2016), for instance, were tiny in size (e.g., a 32x32 pixel resolution image of a “green school bus”). Today, text-guided synthesis of images has made a giant leap towards becoming a mainstream phenomenon (Olson, 2022). Within less than a year, Midjourney’s Discord community has grown to over 10 million users, making Midjourney the largest Discord community to date.[2] Besides more powerful graphics processing units (GPUs), a few particular important inventions advanced the field of text-to-image generation. This section gives a brief overview aiming to explain the recent technical developments that enabled and fueled the meteoric rise of text-to-image generation.

The invention of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) was a watershed moment in advancing image generation. GANs are a type of deep learning architecture consisting of two antagonistic parts: a generator and discriminator. During training, the generator presents the discriminator with synthetic images. The discriminator judges these images and the process is iteratively continued until the discriminator cannot tell the synthetic images apart from real images, such as the images in the training data. Using a text-conditioned GAN architecture, Reed et al. (2016) pioneered the approach of synthesizing images from text. The approach was extended in January 2021 with OpenAI’s DALL-E (Ramesh et al., 2022), a neural network trained on text-image pairs. DALL-E was able to synthesize images from text captions for a wide range of concepts expressible in natural language. In parallel, OpenAI presented CLIP (Radford et al., 2021), a contrastive language-vision model originally conceived for the task of classifying images. CLIP was trained on a large corpus of pairs of images and text scraped from the World Wide Web. Due to the large

[2] See https://discord.com/servers.

size of its training data, the CLIP model has learned a wide variety of visual concepts from natural language supervision. This proved useful for tasks that visually associate language with images. The CLIP model and its training corpus were, however, not released by OpenAI, which spurred efforts to replicate CLIP and its training data.

It was the release of the weights of CLIP in January 2021 that resulted in immense technical progress in the field of AI-generated imagery. The CLIP weights found their first significant application in an image generation system called "The Big Sleep" by Ryan Murdoch (Colton, Smith, Berns, Murdock, & Cook, 2021; Murdock & Wang, 2021). In Murdoch's architecture, the generator is a model called BigGAN and CLIP is used to guide the generation process with text. This inspired Katherine Crowson to connect a more powerful neural network (VQGAN) with CLIP (Crowson et al., 2022). The VQGAN–CLIP architecture became very popular in 2021 and instrumental to advancing the emerging field of text-to-image generation (Crowson et al., 2022). The source code of VQGAN–CLIP was available online, and many generative architectures for synthesizing digital images and artworks have since been developed based on the work by Murdoch and Crowson. GANs were later superseded by diffusion-based systems (Dhariwal & Nichol, 2021). Diffusion models are a class of machine learning models that learn through the introduction of incremental noise into the training data, with the objective of subsequently reversing the noising process and restoring the original image. Once trained, these models are capable of utilizing the learned denoising methods to synthesize novel, noise-free images from random input.

Today, practitioners can choose from a large variety of diffusion-based generative systems. Some of these systems are available as open source, such as StableDiffusion (Rombach et al., 2021), others are available as online services, such as Midjourney and DALL-E 2. Due to the low barrier of entry and high ease of use, Colab notebooks contributed to a democratization of digital art production. Anybody can create digital images and artworks with text-to-image generation systems (Oppenlaender, 2022a), which establishes parallels of the novel technology with photography.