Analysis of Text-to-Image AI Generators

Ziyu Huang (Cheryl)
IPHS300 AI for the Humanities (Spring 2022) Prof Elkins and Chun, Kenyon College

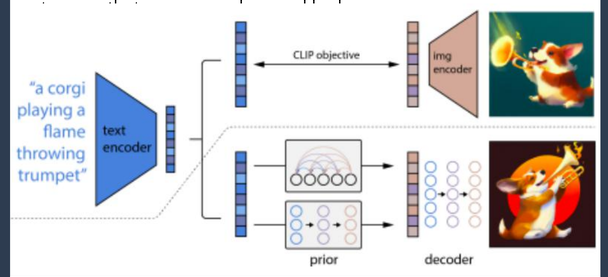
Abstract

This project is an analysis of text-to-image artificial intelligence generators. The comparison will mostly focus on the newly-released DALL-E 2, but will also include two other AI art producers from earlier generations. Each AI generator will be fed the same text prompt for the analysis. Three metrics will be used to analyze the images generated by each AI generator in response to the same text prompt. This project will utilize three metrics: aesthetic, comprehension and interpretation, and creativity. This project will result in a conclusion and a recommendation for the improvement of future AI art generators based on a comparison of the performance of several AI art generators and different text prompts.

Introduction

With the spring 2022 release of DALL-E 2, there is heightened interest in the debate of AI-generated art. In comparison to existing AI art generators that convert text to images, the revolutionary DALL-E 2 is an AI system that can generate more realistic and accurate images based on the text input. Furthermore, DALL-E 2 can make complex artworks with only relatively brief text inputs. In addition to these, DALL-E 2 is capable of visually integrating distinct and irrelevant objects. While earlier AI generators could only produce crude and low-quality images, DALL-E 2 has reached the State of the Art (SOTA) since its products satisfy practically all artistic requirements.

Compared to Generative Adversarial Networks-based model (GAN), DALLE-2 is a newer model that supplants and even excels GAN. Unlike other elementary models that rely mostly on GAN, DALL-E 2 benefits from Contrastive Learning-Image Pre-training (CLIP) and diffusion models. The CLIP parallels the trainings of the texts and images, functions like the encoder; while the diffusion models learn to generate image by nosing and denosing the training set, function like the decoder. DALL-E 2's architecture is to first train the CLIP model and then use it to train the diffusion models. Last but not least, the diffusion models use CLIP to construct text embeddings and generate images corresponding to the text. The most notable benefit of this design is that it does not require massive amount of text-image paired data for training. In other words, it is a model that is unsupervised or "self-supervised." The self-supervised system can save a substantial amount of human labor. At the same time, the unsupervised construct maximizes creativity and novelty, as the AI may discover surprising



Material and Methodology

Material:

1. Twitter posts of DALL-E 2:

As I currently have no access to DALL-E 2, the only sources that can be drawn from DALL-E 2 are from Twitter. This project will therefore collect artwork created by DALL-E 2 from Twitter posts. The spectrum is limited to the arts and excludes photographs. In addition, I will utilize the identical text prompts to feed the other two AI art generators and evaluate the performance of the different generators by comparing their outputs.

2. Hotpot AI Art Maker & 3. Starryai AI Art Generator (Orion)

These are open-sourced AI art generators, featuring fast generating speed (1-2 min) and superior visual quality than other open-source AI art generators.

Methodology:

There is no access to the code underlying these models, thus all evaluation will be based on text input and output images.

All of the text prompts will include an indication of a certain art style and at least one from the subject identification and activity description.

The three metrics developed for this project are aesthetic, comprehension and interpretation(C&I), and creativity. The aesthetic will be the formal analysis of the images produced from the perspective of human art historians. Composition, color palette, and lines and shapes will be the primary factors for conducting the formal analysis. The comprehension and interpretation metric will assess the accuracy with which you comprehend and interpret the text prompt in terms of artistic style, subject matter, and iconography. The creativity will investigate the originality of combining the formal components of the particular art style with the narrative and iconography.

Acknowledgement

Dickson, Ben. "Dall-e 2, the Future of AI Research, and OpenAI's Business Model." TechTalks, April 11, 2022. https://bdtechtalks.com/2022/04/11/openai-dall-e-2/.

O'Connor, Ryan. "How Dall-e 2 Actually Works." AssemblyAI Blog. AssemblyAI Blog, April 22, 2022. https://www.assemblyai.com/blog/how-dall-e-2-actually-works/.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu and Mark Chen. "Hierarchical Text-Conditional Image Generation with CLIP Latents." ArXiv abs/2204.06125 (2022): n. pag.

Swimmer963. "What Dall-e 2 Can and Cannot Do." LessWrong, May 1, 2022. https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do.

Wang, Zihao, Wei Liu, Qian He, Xin-ru Wu and Zili Yi. "CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP." ArXiv abs/2203.00386 (2022): n. pag.

https://twitter.com/Merzmensch/status/1522277446980091904
https://twitter.com/bakztfuture/status/1517373091034378241
https://twitter.com/Merzmensch/status/1523302450047893506
https://twitter.com/Dalle2Pics/status/1521217219488894977/photo/1
https://twitter.com/Merzmensch/status/1523550836281937921/photo/1

Results

Comparison with AI Art Generators from Earlier Generation:

- post-modern style:
"Remembrance of nostalgia, surrealist painting by Dali."

Hotpot:	Starryai:	DALL-E 2:
Aesthetic: 6/10	Aesthetic: 8/10	Aesthetic: 9/10
C&I: 3/10	C&I: 6/10	C&I: 9/10 (closest to Dalí)
Creativity: 5/10	Creativity: 6/10	Creativity: 7/10

- pre-modern style:
"a hot dog in the style of a renaissance painting."

Hotpot:	Starryai:	DALL-E 2:
Aesthetic: 2/10	Aesthetic: 6/10	Aesthetic: 9/10
C&I: 2/10	C&I: 6/10 ()	C&I: 7/10 (more Baroque)
Creativity: 3/10	Creativity: 5/10	Creativity: 6/10

Comparison with Different Text Prompts Using DALL-E 2:

- in the style of Vermeer:
- text prompts from left to right:
"Ai generated 'Robot girl with a pearl earring' by Johannes Vermeer"
"Mother, by Vermeer"
"Good morning, in the style of Vermeer"

Aesthetic: 8/10	Aesthetic: 9/10	Aesthetic: 8/10
C&I: 8/10	C&I: 7/10	C&I: 9/10
Creativity: 9/10	Creativity: 9/10	Creativity: 7/10

- DALL-E 2 generates art by combining the most distinctive and recognizable features of the subject and the style. These "features" may include facial characteristics, costumes, hairstyles, makeup, accessories, color palettes, brushstrokes, modeling of light and shadow, compositions, lines and shapes, etc. But here comes the question, how does DALL-E 2 choose which feature(s) to combine? When text prompts include the name of the style (or the artist's last name if the style is named after the artist), DALL-E 2 is more likely to select the formal stylistic features. In the case above, when "Vermeer" appears as a style, DALL-E 2 generates work with Vermeer's distinctive sketchy brushstrokes and bluish, cold-toned color palette. While the first does not incorporate Vermeer's painting style.

Conclusion and Recommendation

Comparing the performance of AI art generators, DALLE-2 outperforms earlier generations of AI art generators on all three metrics. It can generate images with a high level of aesthetic quality, an accurate interpretation of the text prompt, and some creativity in blending information with style. Nonetheless, the outcome demonstrates that DALL-E 2 has several limitations. First, its spelling ability is relatively poor. When asked to generate graphics with some text on them, typographical errors are quite probable. Several DALL-E 2 users are also aware of this shortcoming. Second, it has different levels of art style comprehension. It has a greater understanding of postmodern and contemporary art styles, especially digital art and some cartoon styles linked to popular animations. According to one of the user reports, DALL-E 2 has trouble assigning specific attributes to particular characters. This circumstance occurs when the text prompts involve two or more figures and indicate distinct characteristics for each figure. In addition to some fundamental characteristics like as gender, DALL-E 2 can easily mix up age, hairstyle/color, and clothing. Even while DALL-E 2 exhibits its strength in analyzing and comprehending subjects, it cannot create satisfactory results when the text prompt contains a novel subject, as stated in the same user report.

The majority of these constraints can be overcome by by modifying the parameters of the DALL-E 2 model. For example, the disparity between the amounts of accessible digital data for works of art generated throughout different eras is the primary cause of different degrees of comprehension of art styles. The majority of the premodern artworks are paintings or sculptures on easels. Their reliance on artistic expertise and lengthy production time restricts their quantity, and many of them are damaged or destroyed. Postmodern artworks, in this case the digital arts, require less painting or sculpting expertise and less time to execute. Therefore, there is a disparity in the amounts of artworks created throughout different time periods, which persists in the DALL-E 2 training data. This bias in the trainning data results in various levels of art style comprehension. However, this could be improved by altering the parameter to have more pre-modern iterations than post-modern iterations.

Currently, there are numerous critiques about the ethical issues posed by Deepfakes created by AI art generators. However, as several users have pointed out, DALL-E 2 appears to have deliberate flaws in its ability to generate photorealistic human faces. Some say that this flaw is one of DALL-E 2's defects. However, DALL-E 2 is capable of producing photorealistic images of objects and non-human animals. Therefore, it is more plausible to believe that the flaw is an intentional attempt to prevent the creation of Deepfakes. One of the additional worries regarding DALL-E 2 is that the AI art generators may lead to the unemployment of artists, particularly digital artists. DALL-E 2's exceptional s creativity can occasionally surpass human intelligence, as it can produce combinations of style and content that have never been observed by humans. However, rather of eliminating employment, AI art producers are more likely to change them. For instance, AI art generators like DALL-E 2 requires domain expertise to improve the perfomance.