Create New Artworks Using Generative Deep Learning with Disco Diffusion

Github Repository:

https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject

OVERVIEW

This experimental project uses Disco Diffusion to generate artworks by experimenting with different text prompt and parameters to understand how CLIP Text-to-image generation works. Disco Diffusion is a Google Colab Notebook that leverages an AI Image generating technique called CLIP-Guided Diffusion to allow you to create compelling and beautiful images from just text inputs. First, I used the default setting to understand how it works. After that, I tested four more text prompts to instruct the AI to generate images that are hopefully very close to what I am picturing. In the end, I generated two videos with different text prompts and parameters based on my understanding of the results from the previous experimental images. As a result, it successfully created a stunning and unexpected video from Disco Diffusion. The following will describe the experiment process and explain how CLIP Text-to-image generation works.

Started with the default setting

The image was created with the default setting in Disco Diffusion using the text prompt: "A beautiful painting of a singular lighthouse, shining its light across a tumultuous sea of blood by greg rutkowski and thomas kinkade, Trending on artstation."

Diffusion is an iterative process. Each iteration, or step, CLIP will evaluate the existing image against the prompt and provide a direction to the diffusion process. Diffusion will denoise the existing image, and it will display its current estimate of what the final image would look like. Initially, the image is just a blurry mess, but as Disco Diffusion advances through the iteration timesteps, coarse and then fine details of the image will emerge.

The image with the default text prompt took 250 diffusion steps to complete. As you can see in the image sequence above, the images get progressively clearer over the range of steps, as the diffusion denoising process is guided toward the desired image by CLIP.

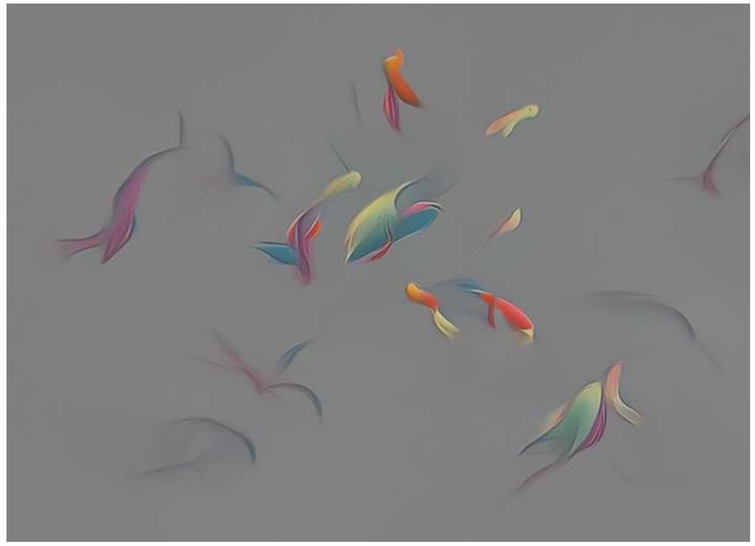
Steps 1, 50, 100, 150, and 200 of the diffusion process

"A beautiful painting of a singular lighthouse, shining its light across a tumultuous sea of blood by greg rutkowski and thomas kinkade, Trending on artstation."

1st Attempts at prompting art

The main idea is to describe what I would now say is more a search term than an imaginative prompt, then the neural network will attempt to find images and merge or evolve them into what it is you are trying to describe. The image was the very first one I tried, and I have to say I was expecting something a bit more colorful; instead, I just got a blurry goldfish.

50% of the diffusion process



"*A colorful pattern of goldfish, Trending on artstation.*"

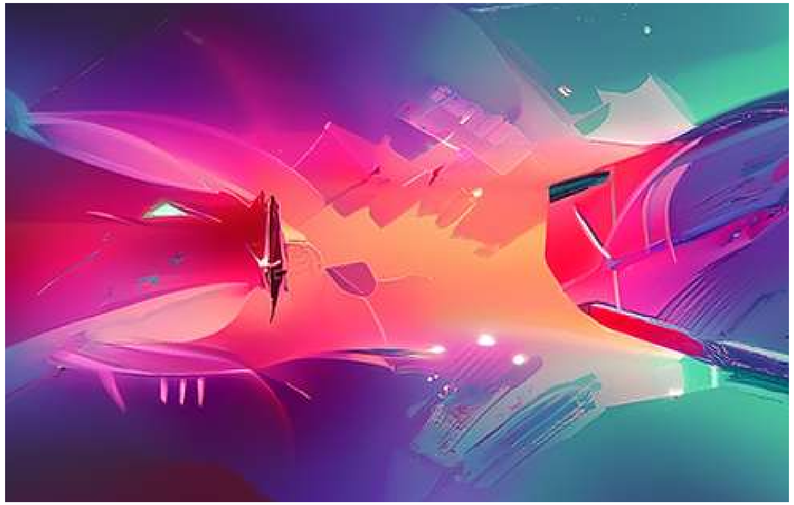
Github link: https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/Images

2nd: Experiment on different prompts in the same image

For this, I want to know how's the AI identify different text prompt in the same image. As you can see from the result, both images don't have an exact outline shape of spaceship or self-portrait, because the text prompt is not clear enough.

Original Image

Text Prompt 1: "A beautiful spaceship, Trending on artstation"



Text Prompt 2: "AI self-portrait"

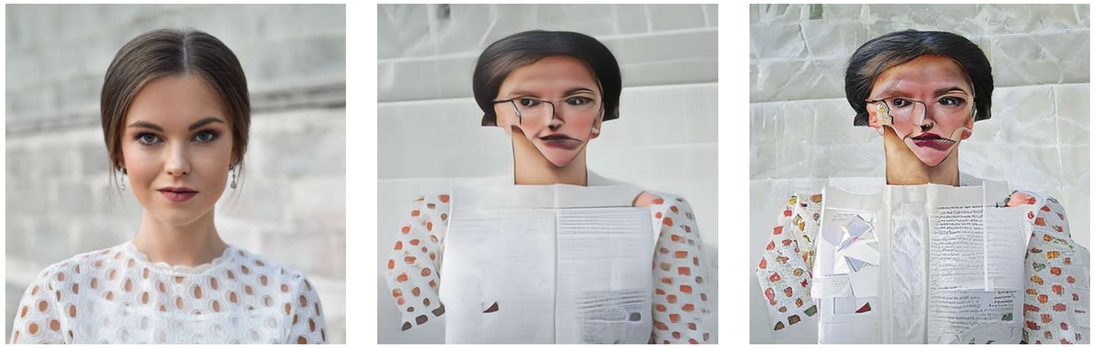
Github link: https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/Images

3rd: Experiment on adding details in the text prompts

According to the examples in Disco Diffusion, text prompt loosely follows a structure: [subject], [prepositional details], [setting], [meta modifiers and artist]. I tried to be rather specific in this text prompt and then add what I call modifiers after the main instruction. For image one, it looks better than the previously generated images, and it does fit the text with "paper collage". Another image added a style text prompt, just changing it a bit with playful and colourful style, then it can fulfil a style and looks more complete, still abstract and even just nonsensical but cool nonetheless.

According to the examples in Disco Diffusion, text prompt loosely follows a structure: [subject], [prepositional details], [setting], [meta modifiers and artist]. I tried to be rather specific in this text prompt and then add what I call modifiers after the main instruction. For image one, it looks better than the previously generated images, and it does fit the text with "paper collage". Another image added a style text prompt, just changing it a bit with playful and colourful style, then it can fulfil a style and looks more complete, still abstract and even just nonsensical but cool nonetheless.

"A paper collage and realism style portrait"

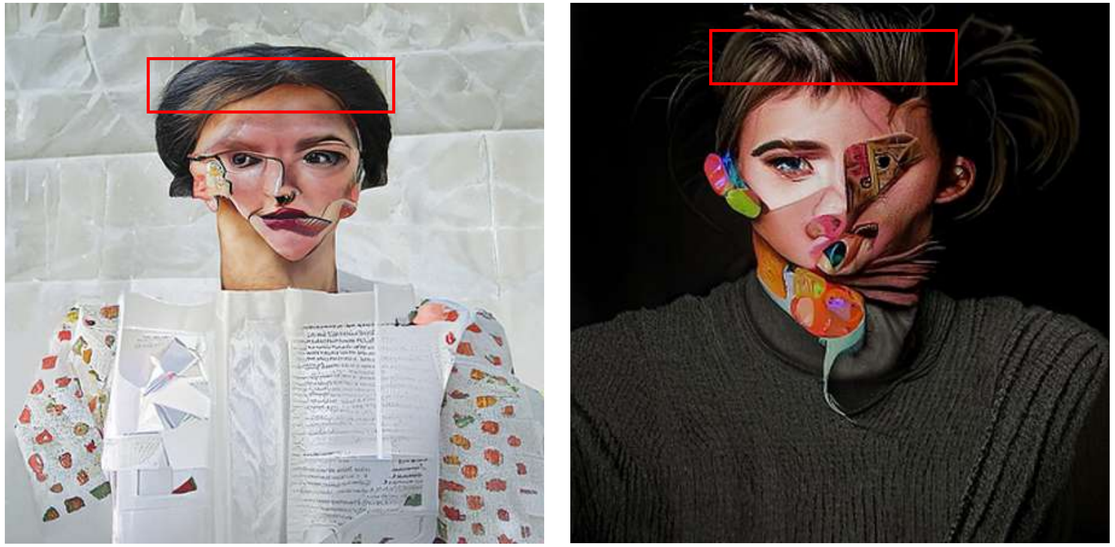


"A paper collage and realism style portrait", "colorful scheme", "Playful"

Github link: https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/Images

Different Parameters on Step

Besides the above text-prompt texting, I also try to use different parameters on Step. Each step (or iteration) involves the AI looking at subsets of the image called 'cuts' and calculating the 'direction' the image should be guided to be more like the prompt. Increasing steps will provide more opportunities for the AI to adjust the image, and each adjustment will be smaller, thus yielding a more precise, detailed image. However, it also will increase the rendering time. For the 3rd experiment on image two, I increased the step to 1000. The result looks more detailed on hair compared to image 1. However, if using an init_image, need to skip ~50% of the diffusion steps to retain the shapes in the original init image. I didn't test the parameters lower than 50% as I don't have enough time to run more testing.



Github link: https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/Images

1st 2D ANIMATION

The video was created with a basic setting - two text-prompt and 120 frames. To generate a video, need to define the parameters on 'angle', 'zoom', 'translation_x' and 'translation_y'. For the angle parameters, I changed to "0:(3.0), 12:(3.0), 24:(0)" as I want to rotate the image by 3.0 degrees from the beginning and in frame 12. Moreover, zoom values over 1.0 are scale increases. Thus frames 24, 36 and 48 will zoom into an image. Cerama view also edited in 'translation_x' and 'translation_y' to shift the image to the right and left. However, the init image is quite complicated, and I didn't change the skip step parameter to lower. It causes the outcome can't see the fairy and yellow colour scheme' distinctly.

Text-Prompt

"0": ["A beautiful painting of a singular castle, shining its light across a tumultuous sea of flower, aesthetic and illustration style.", "yellow color scheme"],
"60": ["A beautiful painting of a singular castle surrounding by fairy, aesthetic and illustration style.", "pink color scheme"]

Github link:

https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/2DAnimations

Youtube: https://youtu.be/GGmDCZpX3z8

LEA HO (20035099), 17 March 2022

2nd 2D ANIMATION

I added more text prompts in different frames for the second video to make the video more smooth and more exciting. Compared with the first video, this video looks more fluent and obvious that images can be linked to text prompts.

Text-Prompt

"0": ["museum books library magical"],
"10": ["museum books library magical | bubble sphere artstation"],
"40": ["museum books library magical | phoenix bird illustration artstation art picture flaming depiction mythology hell"],
"50": ["museum books library magical | unicorn illustration artstation art picture flaming depiction mythology"],
"70": ["museum books library magical | artists concept artstation star megastructure universe space orbit sphere"],
"78": ["museum books library magical | artstation past nature wonderland],
"90": ["museum books library magical | building disneyland artstation"],
"100": ["museum books library magical | walking around artstation rain"],
"107": ["museum books library magical | artstation crowded street people aerial view"],
"133": ["museum books library magical | working crowds artstation aerial"],
"148": ["museum books library magical | happy people walking crowd"],
"157": ["museum books library magical | rainy day bookshop book store artstation"],
"179": ["museum books library magical | pink glow rain street artstation"],
"200": ["museum books library magical | life in the past old pictures"]

Github link:

https://github.com/leahoho/CCI_CodingTwo/tree/main/FinalProject/DiscoDiffusion/2DAnimations

Youtube: https://youtu.be/OwNfKid53U4

To sum up, the main event to generate video or image with Disco Diffusion is the text-prompt and tune the parameters. Although fine-tuning the prompt and parameters is complex and time-consuming, it really helps to understand more how AI generate artwork. CLIP really depends a lot on specific keywords, probably due to its dataset using a major image hosting website. Therefore, a significant part of generating art is optimizing your text prompts (Prompt engineering).

In addiition, a big thing that confused me when using the workbook and rendering was that sometimes it would take up to 3 hours to render only one image. For video, it takes more than 6 hours to render 120 frames. So you won't know the outcome if it meets your expectations. And Disco Diffusion has dozens of controls, with complex interactions and few limits, so it's easy to get bad results. Furthermore, I didn't have the opportunity to experiment more because of the GPU limitation of Google Colab, and it limited me to creating more videos to test with different interaction settings.