A Close Encounter

Damian Franco
CS 429/529 Introduction to Machine Learning

Description

My artwork was inspired by my upbringing in the city of Roswell, New Mexico. If you ever visit Roswell, you cannot escape the story of an alien crash landing that allegedly occurred there in 1947 and with the countless alien decorations around the city. As a child, my family and I would go outside and stargaze, hoping to see the unexplainable. Needless to say, there were things that I could not explain that I have seen when stargazing.
The artwork itself depicts an object in the night sky that could resemble something from out of the world. I personally wanted the point of view from the artwork to be from the ground to give it the effect that it is an individual that is viewing this object in the night sky. The image just needed a bit more of a city feel, so I also prompted the model to generate some buildings in the image to give it more perspective. Overall, I wanted to image to have the feeling that I felt when growing up and stargazing and having the chance of seeing something unexplainable.

Craiyon

Craiyon, formally known as DALL-E Mini, is a variant of OpenAI's DALL-E image generation model developed by Boris Dayma, et al. While the original DALL-E model can generate high-quality images from textual descriptions, Craiyon is designed to do the same but it is more computationally efficient version that can run on devices with limited computing resources, such as mobile phones or embedded systems. [1]
Craiyon uses a transformer-based architecture, similar to the original DALL-E model and other state-of-the-art language and image models. It takes a textual prompt as input and generates an image that corresponds to the prompt. The prompts can be simple phrases or complex sentences describing an object or a scene. To generate the artwork above, I used the prompt "ufo from the ground shot in black and white with rural buildings in the lower view".

Techniques

The model works by encoding the textual prompt into a vector representation and then decoding it into an image. During the decoding process, Craiyon generates a sequence of pixels that make up the final image. The model is trained on a large data set of image-caption pairs, which allows it to learn the patterns and features that correspond to different objects, colors, and textures. Images are encoded through a Vector Quantized Generative Adversarial Network or VQGAN encoder, which turns images into a sequence of tokens. Descriptions are encoded through a BART encoder. The output of the BART encoder and encoded images are fed through the BART decoder, which is an auto-regressive model whose goal is to predict the next token. Loss is the softmax cross-entropy between the model prediction logits and the actual image encodings from the VQGAN. [2]
More specifically, Craiyon follows a the following process. First, the textual caption is encoded using the BART encoder. Next, a special "Beginning Of Sequence" token is fed through the BART decoder, which then samples image tokens based on its predicted distribution over the next token. These sequences of image tokens are then decoded through the VQGAN decoder. Finally, the CLIP model is used to select the best generated images based on their similarity to the original textual caption.
Craiyon is a powerful tool for generating images from textual descriptions in real-time. It has a wide range of potential applications and generated the great piece of art, alongside many more that I wish I could show as well.

References

[1] Dayma, B, Patil, S, Cuenca, P, Saifullah, K, Abraham, T, Le, P, Luke, Ghosh, R. (2022). *DALL-E Mini – Generate Images From Any Text Prompt*.
[2] Dayma, B, Patil, S, Cuenca, P, Saifullah, K, Abraham, T, Le, P, Luke, Ghosh, R. (2022). *DALL-E Mini Explained*. https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-mini–Vmlldzo4NjIxODA