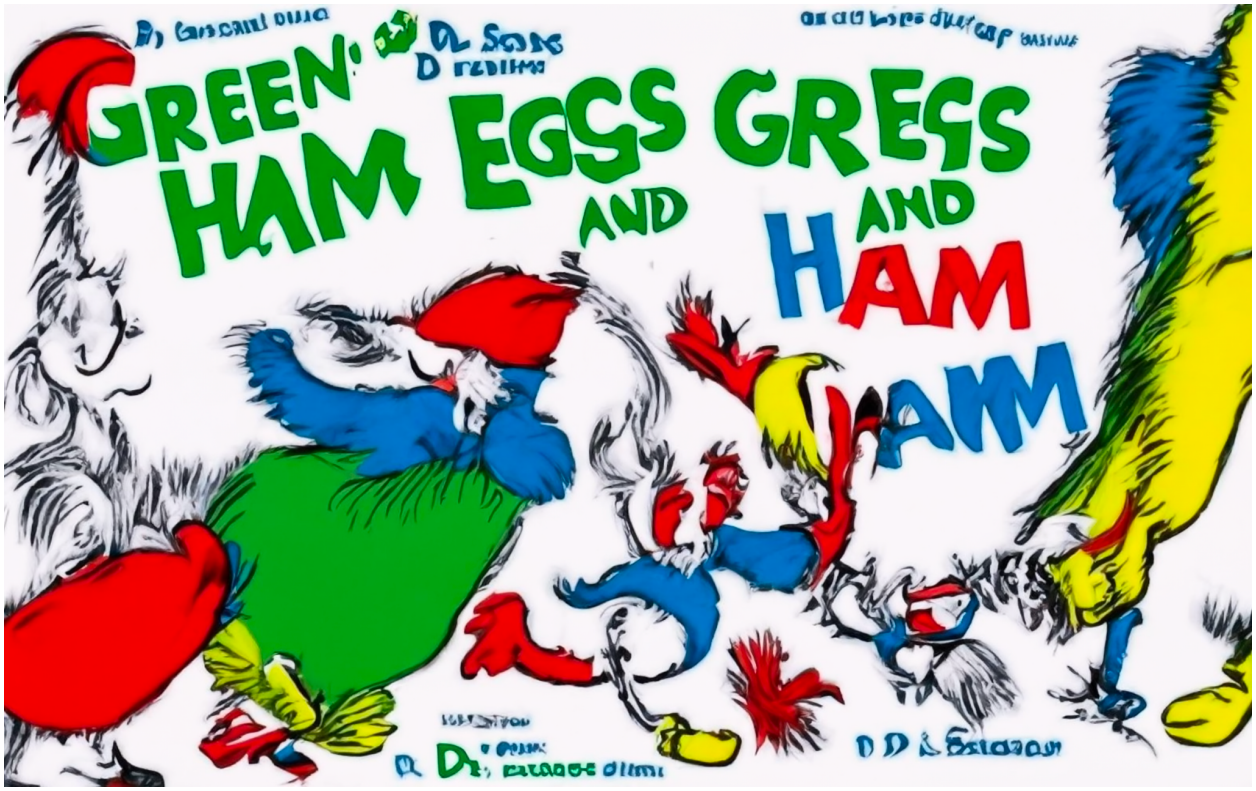
ART AND MACHINE LEARNING
CMU 2022 SPRING
PROJECT 4 GROUP 18

Children's Stories to give you Nightmares



Francisco Cabrera (School of Computer Science, Bachelors)
Xiwen Chen (Language Technologies Institute, Masters)
Sebastian Montiel (Electrical and Computer Engineering, Masters)

DESCRIPTION

In this project we used a text generator to generate a short story using a prompt. Then we used image generating methods to create images with text prompts that match the sentences in the generated text. We experimented with different image generation techniques and included multiple styles for each generated story. The final results make up a book with machine generated child stories with illustrations in different styles.

Concept

Inspired by GPT-3 which achieved great results in text generation tasks, we aim at machine generated stories with illustrations. With the first few sentences from the original story from famous child story authors/illustrators, we first generated the complete story with a state-of-the-art neural network, and then created illustrations from these generated texts using image generation techniques.

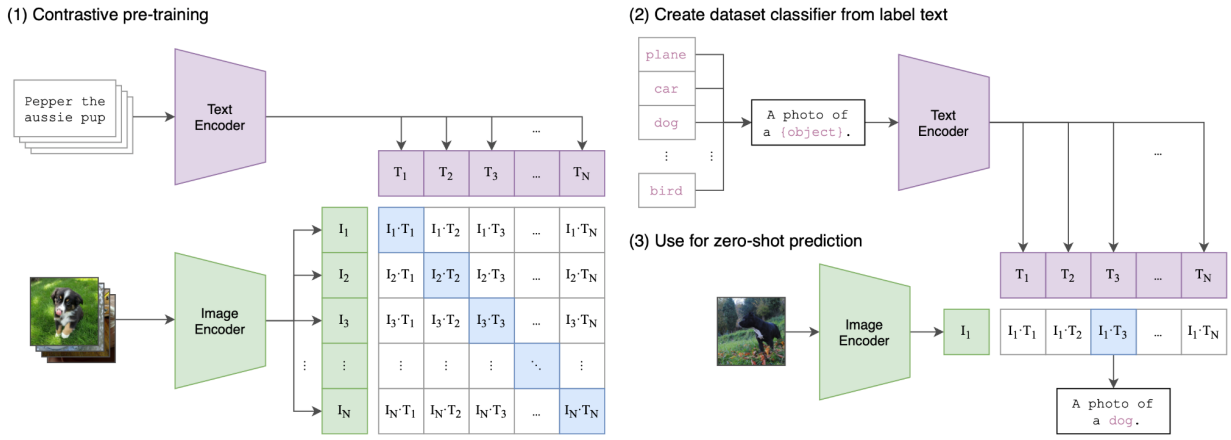
Technique

Text Generation

There exist various text generation techniques including GPT models and transformer based models. In order to generate texts that match the style of a story, we used a text generating model that is trained on web texts. Specifically, we used InferKit to generate at most the next 1000 words from a few sentences of prompt. The prompts are found from child stories available online.

VQGAN-CLIP

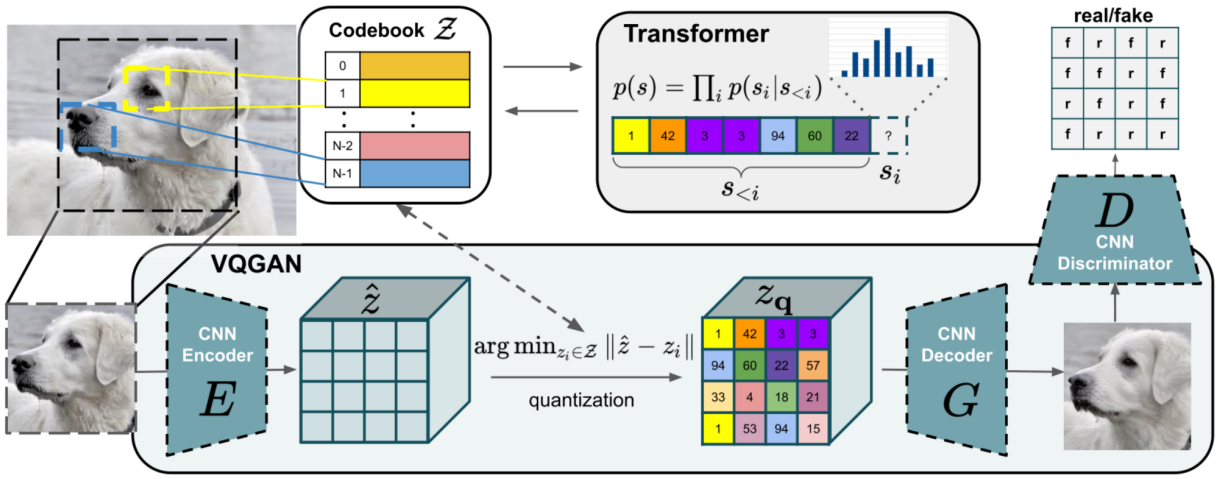
Many multimodal models are able to generate images with the guidance of text prompts. They vary in styles and can also include abstract objects described in the prompt. For the first attempt, we used VQGAN to generate images for each sentence of the generated texts. The VQGAN model is based on CLIP from OpenAI, a visual model with natural language supervision.



CLIP model.

As is shown above, the CLIP model contains a text encoder and an image encoder, and is trained to match the correct pairings of image and text. The VQGAN-CLIP model uses this architecture and trains

the encoders and decoders using VQGAN. The model is then able to create images from text prompts in artistic style.



VQGAN

Image Generation with Latent Diffusion

Latent diffusion uses diffusion models in the latent space of pretrained autoencoders and cross-attention layers to use the diffusion models as generators. We used a latent diffusion model trained on the LAION-400-Million dataset consisting of image text pairs to facilitate generation of images from text.

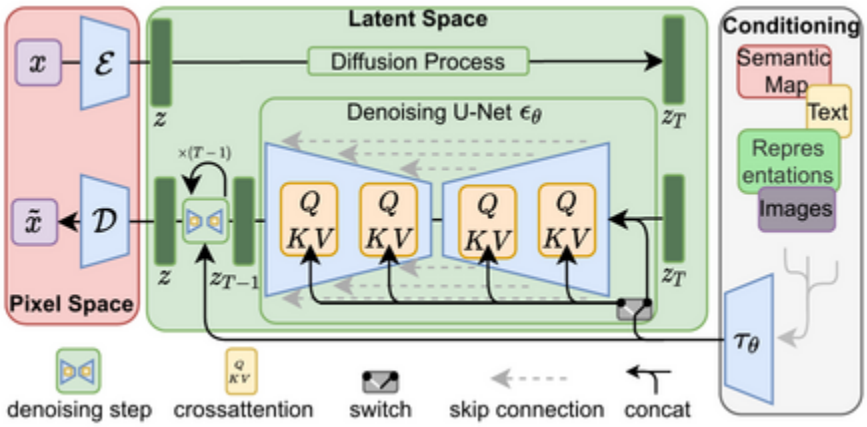


Image Generation with CLIPDraw

CLIPDraw is an algorithm to generate drawing based on text input using a pre-trained CLIP model to optimize the similarity of the drawing and text input description. We discussed this algorithm in class so we'll spare the details here.

Process

Text Generation (Xiwen)

We used InferKit to generate the next 1000 words using the first few sentences from child stories available online. An example output is shown below.

Prompt:
Once upon a time there was a dear little girl who was loved by everyone who looked at her, but most of all by her grandmother, and there was nothing that she would not have given to the child.

Output:
When the winter came the grandmother set aside the rich food to make a sweater, and before the girl went to school, a walk was arranged, and afterwards there was a kiss, which was longed for by both.

The girl loved her grandmother, and now and then took some food to her.

Her grandmother lived alone and was only able to dress herself, and had never heard anyone call her "mother."

One morning she was shaving, the door of her room was opened by a very little boy, and to her astonishment she saw that it was her own dear little darkey.

He had managed to break into the room and was standing before her. "Mother," said the boy, "I was obliged to come here this morning.

I know where you live, and I have been here three days, and I have found nothing to eat, and my nails are all cut, and I don't think I shall be able to go home any more.

I have asked the others, but nobody seems to want me.
Are there no people in the world who would want me?
Do you know who I am?"

As we see from the results, the model is able to generate stories that might come along with the main characters mentioned in the prompt texts. The generated texts are generally semantically correct, and follow the style of the prompt.

Image Generation with VQGAN-CLIP (Xiwen)

We then used VQGAN-CLIP to generate illustrations for the generated stories. Using each tokenized sentence from the stories as input, the model then generates images that include features indicated in the sentence in artistic style. An example of outputs is shown below.

There was once a velveteen rabbit, and in the beginning he was really splendid.

He was fat and bunchy, as a rabbit should be; his coat was spotted brown and white, he had real thread whiskers, and his ears twitched by themselves.

On Christmas morning, when he sat wedged in the top of the Boy's stocking, with a sprig of holly between his paws, the effect was charming.

As we can observe from the results, the model is able to integrate the features, “rabbit” for example, into the output image. The images are in painting styles with content summarized from the text. Therefore, they could serve as illustrations in the created stories.

Image Generation with Latent Diffusion (Francisco, Sebastian)

We used a few different techniques to generate images for the stories using Latent Diffusion. One technique we used was just prompting with text directly from the stories. Another technique tried was prompting with things along the lines of “Detailed book illustration by Dr. Seuss of: …” Finally, we tried prompting with a more vague style like “Watercolor painting of …”. Some results are shown below.

Story	Prompt	Output
Green Eggs and Ham	Watercolor painting of green eggs and ham	

The Little Red Hen	Watercolor painting of a red hen in a barnyard	

Image Generation with CLIPDraw (Sebastian)

Similar to the process used for latent diffusion image generation, we used descriptions of things occurring in the stories to generate corresponding images that could be combined into a children's picture book. Some sample outputs are shown below.

Story	Prompt	Output
Green Eggs and Ham	Watercolor painting of green eggs and ham	
The Little Red Hen	Watercolor painting of a red hen in a barnyard	

Centipede Diffusion

Centipede Diffusion takes a Latent Diffusion output, upscales it, and uses it as input to a disco diffusion model. This creates a more artistic style since the disco diffusion model can focus more on details rather than a correctly classified image.

Reflection

In this project we applied text generator and image generator to create a children's book with illustrations. We experimented with different image generation techniques, and observed that many of them are able to paint like drawings with objects and features mentioned in the text prompt, and sometimes even including some texts in the image. In general, we have successfully created a book with illustrations as our goal suggests, and this could be inspiring for future research in machine created artworks.

RESULT

CLIPDraw and Latent Diffusion Stories:
https://docs.google.com/presentation/d/1VZfDxOanqIOuGi-GYccTTJkM_aUOoiYIZxVlxcGycoA/edit?usp=sharing

Book of stories from multiple image models:
https://docs.google.com/document/d/1gN5aItTet7C61PGEXTTBD2PKJosnHiuTImCjrwLIako/edit#heading=h.3o1ocxea3cws

CODE

VQGAN-CLIP
https://github.com/chxw20/vqgan-clip

Latent Diffusion
https://colab.research.google.com/github/multimodalart/latent-diffusion-notebook/blob/main/Latent_Diffusion_LAION_400M_model_text_to_image.ipynb?authuser=1#scrollTo=NUmmV5ZvrPbP&uniqifier=1

CLIPDraw
https://colab.research.google.com/github/kvfrans/clipdraw/blob/main/clipdraw.ipynb

REFERENCE

[1] Esser, Patrick, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.