Mosaic: Generating The New Yorker Style Cartoons using Text-to-Image Diffusion Models

Mehul Sudrik, Rajesh Nagula, Utsav Oza

ECE-GY 7123 Deep Learning

Problem Statement

> "Sometimes I wonder if I'm too old to be a cartoonist, but then I remember that I'm just not funny enough."

This self-deprecating joke highlights the challenges of creating The New Yorker style cartoons, that uniquely combines a whimsical art style, witty humor, and a subtle commentary on modern life. Cartoon enthusiasts and creative professionals alike know that creating such cartoons can be a daunting task that requires both artistic talent and a knack for satire. In this project, we aim to explore techniques to simplify the cartoon creation process by using Text-to-Image Diffusion models to specifically generate high-quality The New Yorker style cartoons from natural language captions.

Literature Survey

Text-to-image diffusion models are a type of generative model that can produce high-quality images from textual descriptions. Diffusion (Ho, Jain, and Abbeel 2020) is a probabilistic process that involves gradually adding noise to an image until it becomes completely random, and then gradually removing the noise until it becomes the target image. The goal of diffusion models is to learn the latent structure of a dataset by modeling the way in which data points diffuse through the latent space.

The challenge, however, is that it is unclear how the diffusion process can be directly exercised to generate images of specific unique concepts, or compose them in new roles and novel scenes. Textual Inversion (Gal et al. 2022) allows us to teach text-to-image diffusion models new concepts - it takes a very small number of images of a user-provided concept, like an object or a style, and learns to represent it through new "words" in the embedding space of a frozen text-to-image model. These "words" can be composed into natural language sentences, guiding personalized creation in an intuitive way.

Project Considerations

Dataset

A key challenge in fine-tuning a text-to-image diffusion model is the requirement of a large and diverse dataset, which can be difficult to obtain for a specific domain, especially if it involves rare or complex visual concepts. Textual Inversion allows us to teach an image generator a specific visual concept through the use of fine-tuning using very few image examples. As such, we'll be relying primarily on the dataset derived for The New Yorker's Cartoon Captioning Contest (Hessel et al. 2022), which is composed of raw cartoon images that are mapped to different caption choices, ranked based on quality and an explanation describing the underlying humor of the cartoon.

Model

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb). Consequently, if your file is improperly formatted, it will be returned to you for correction.

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb). Consequently, if your file is improperly formatted, it will be returned to you for correction.

Goals

Please check all the pages of your PDF file. The most commonly forgotten element is the acknowledgements — especially the correct grant number. Authors also commonly forget to add the metadata to the source, use the wrong reference style file, or don't follow the capitalization rules or comma placement for their author-title information properly.

References

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Hessel, J.; Marasović, A.; Hwang, J. D.; Lee, L.; Da, J.; Zellers, R.; Mankoff, R.; and Choi, Y. 2022. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. *arXiv preprint arXiv:2209.06293*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.