Project Task: Text to Movie Poster Generator

Luis Simplicio Ribeiro, Alison Yao, Isidora Diaz

March 24, 2023

1 Summary

We will try to fine-tune the stable diffusion model so that we can input text (eg. movie title, actor names, plot summary) and generate movie posters accordingly. In order to fine tune the model we will use data from IMDB, which contains poster images and movie information that can be retrieved using web scraping.

2 Computational Task Description

The task is conditional image generation, where given a prompt (text input), the model has to generate a movie poster that matches it. The data we are going to use include movie posters and text information on the movies such as movie title, actor names, and plot summary. We might need to engineer the text prompt before inputting into the model so that the model can understand the prompt better. Conditional image generation is a very hard problem in general, and starting from scratch would be very difficult, so we are going to explore using pre-trained models and somehow either fine tune these models or optimize the prompts to generate the images that we are interested in. Additionally, movie posters usually have text on them but these text are not conducive to model training. We will need to eliminate the text information on the movie posters' output by processing our images using `detextify` [1], a python library that removes unwanted pseudo-text from AI generated images.

3 Brief Literature Review

Current state of the art of text to image machine learning models include DALL-E 2 [1] and Stable Diffusion [2]. These models were trained in millions to billions of images.

Diffusion models [3] comprise a two step process, where given data from a target distribution, it destroys structure on data, until it becomes noise, and then it learns how to recover the original input. If the learning process is successful, we can sample noise and obtain new samples from the target data distribution. Stable diffusion [2] conducts the diffusion process in a low dimension latent space, by encoding the original image with an Encoder network, and mapping it back with a Decoder network. Additionally, they are able to generate better images by guiding the generation with text inputs.

In a similar fashion, DALL-E 2 [1] can create realistic images from text descriptions (prompts), with a generative stack consisting also of two stages. In the first stage, a prior generates CLIP image embeddings conditional on their respective prompts, and in the second stage a decoder generates images conditioned on CLIP image embeddings and, optionally, text captions. DALL-E 2 was developed by OpenAI and is not an open source project, for this reason we won't be able to transfer from this model.

[4] gives us some guidance on how we might fine-tune or even re-train large models such as stable diffusion on personal devices. ControlNet manipulates the input conditions so that it can oversee and control the behavior of the entire network and its outputs. We will explore its Github code and see how we can employ ControlNet for movie poster transfer learning.

[1] `https://github.com/iuliaturc/detextify`

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[3] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[4] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.