Project Method: Text to Movie Poster Generator

Luis Simplicio Ribeiro, Alison Yao, Isidora Diaz

April 14, 2023

1 Computational problem

The goal of this project is to generate movie posters using descriptive text prompts. The movie poster data will come from IMDb. Depending on the model, we will explore inputting only text prompts as well as images paired with text prompts. The prompts could include information about the title of the movie, the movie, the actors, the genre, and so on. We will engineer these textual prompts to guide the stable diffusion model to generate movie posters. We will also attempt to fine-tune stable diffusion model using these image-text pairs. We expect that it is easier to generate the posters of existing movies rather than non-existing movies. The goal is to explore what kind of prompt generates better movie posters and how well we can generate variations of movie posters. We anticipate the following:

1. There is little guideline on which prompt works and which does not. We will need a lot of trials and errors to understand the temperament of stable diffusion model.
2. Fine-tuning the stable diffusion model might be challenging for numerous reasons:
   (a) The model has too many parameters, meaning that the training would take too long and would require too much computational resources depending on the size of the training dataset. With a model of this size, even loading it on memory can be problematic.
   (b) Fine-tuning a model in these circumstances could lead to a problem known as catastrophic forgetting, where the model radically forgets information learned in the past, given the new information that it has to learn. Therefore, fine-tuning the model could actually have the effect of degrading its performance.
3. There are not quality evaluation metrics for generated output that are specific for movie posters. We will discuss our approach in the Evaluation section.
4. Movie posters have text on them, but these models were not trained to explicitly generate images with text. Therefore, we should expect that the generated images will contain text that does not even make sense.

2 Choice of method

Formally, our task is conditional image generation: we want to have access to a model $p_\theta(x|y)$, which is able to create a rendition of a realistic movie poster x, whose content is given by a textual description y. In order to solve this, we will use Stable diffusion [8], a model that was trained on million of images-caption pairs (stemming from a subset of LAION-5B [11]). The model is composed of a frozen CLIP ViT-L/14 [6], a 123 million parameters text encoder that map a text

prompt to vector representation. This vector representation is then combined with a sampled noise, and given as input to a 860 million parameters UNet [9], which then generates an output image.

To solve this problem, instead of learning the model $p_\theta(x|y)$ from scratch we are testing two strategies: **Zero-shot learning**, where given that the model was training in a very large dataset of image caption pairs, we do not fine-tune the model, instead we try to find the the prompts y that will lead the model to generate realistic poster images that matches the prompt. With **fine-tuning** we create a small image-prompt dataset and fine-tune the stable diffusion model so it is somewhat biased towards our prompts, by doing the following:

1. We will choose some images of movie posters of selected genres from different studios and label them (create the prompts). The prompts could contain for instance background, the number of characters in the poster, some traits of them such ethnicity and gender, and a short description of the plot of the movie.

2. For an image x_i we will create a set of n slightly different prompts, $y_{ij}, j = 1, \cdots, n$, that matches to that image. Leading to a dataset $D_i = \{(x_i, c_\phi(y_{i1})), \cdots, (x_i, c_\phi(y_{in}))\}$, where c_ϕ is the CLIP model used to encode the prompts.

3. Our final dataset will be composed by the union of the datasets D_i, for each image one of the m images: $D = D_1 \cup \cdots \cup D_m$.

4. We then use the diffusers library [12] to perform training on a Stable Diffusion model.

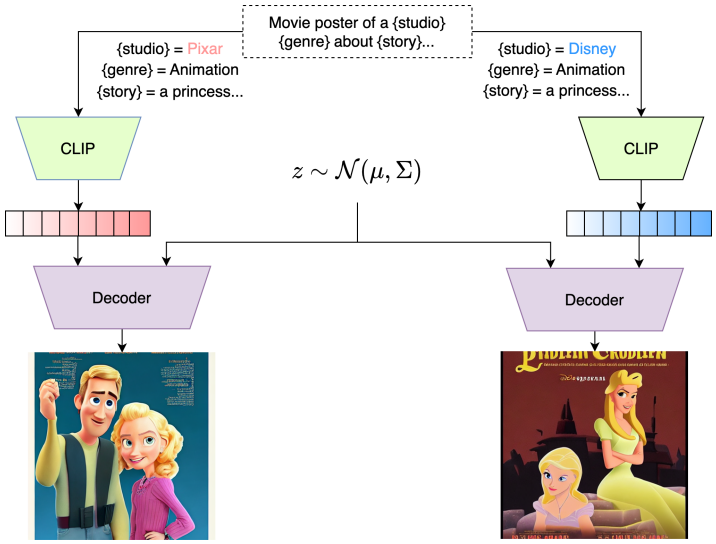

Figure 1: General framework of generating movie posters with Stable Diffusion

3 Match to the problem

We are focusing on generating images conditionally on the text input, and we will utilize a model that was originally intended for a broader task (generating any image from text, as opposed to movie posters in particular).

We propose two methods in the previous section that require minimum computational power, which is in the scope of this class project. Our solution most likely won't be able to remove pseudo-text from the generated images, however we will explore external libraries that have been developed to tackle this problem ex post (such as `detextify`).

4 Alternative Methods

The main component in this task is being able to generate a realistic images conditioned on text prompts. There are several text-to-image generative models available, based on Generative Adversarial Networks [3] and Variational Autoencoders [4] for instance, that we could use, but since recently diffusion models have shown to be a powerful alternative, these are the models we decided to use. For this particular problem we will be using Stable diffusion, which is not only one of the models generating the best quality images, but is also open source, as opposed to models such as Midjourney and DALL-E 2 [7] (which we have access to the paper but not the official implementation by OpenAI). On top of that, it is very lightweight compared to other models.

5 Evaluation

Evaluating the generated posters is not straight forward, however the most common metrics to assess image-from-text samples include automatic image quality and diversity, automatic image-text alignment and human evaluation.

First, for automatic image quality assessment we can use the Fréchet Inception Distance (FID), inputting both generated and real posters through the Inception v3 [13] model, extracting features from the last pooling layer of the model and then using them to fit two separate multi-variate Gaussians. The FID score corresponds to the Fréchet distance between the two multivariate Gaussian distributions [14]. However one limitation of this metric is that it is not aligned with perceptual quality [10]. Alternatively, Inception Score (IS) evaluates the diversity of output images, because we would like the marginal probability of labels over all generated images to be diverse instead of having the same poster generated every time [1].

Second, automatic image-text alignment can be measured by taking the generated images and inputting them into a captioning generator pre-trained model (such as VL-T5 [2]), and then comparing similiarity between both texts using a metric (such as BLEU [5]). Another alternative is to use the CLIP model and input pairs of text and image (caption and poster, real and generated), and then comparing the similarity of these pairs in the latent space.

Finally, given the subjective nature of generative movie posters, it could also be valuable to have humans evaluate pairs of generated and real images on image quality and image-text alignment. The most common strategy is to present evaluators with side-by-side comparisons between generated and real images, and have them choose which they like better and which one resembles better the prompt.

References

[1] Yaniv Benny, Tomer Galanti, Sagie Benaim, and Lior Wolf. Evaluation metrics for conditional image generation. *International Journal of Computer Vision*, 129:1712–1731, 2021.

[2] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[7] 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.

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

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[11] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[12] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[13] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 783–787. IEEE, 2017.

[14] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.