Artist's Assistant: Assist User on Prompt Creation For Text-to-Image Synthesis

Jiayi Pan, Yuanli Zhu, Huanshihong Deng
{jiayipan, leozhu, denghsh}@umich.edu

1 Introduction

(a) Generated with prompt: "San Francisco".

(b) Generated with prompt: "San Francisco, RPG Reference, Oil Painting, Trending on Artstation, octane render, Insanely Detailed, 8k, HD, unreal 5, DAZ, hyperrealistic, dynamic lighting, intricate detail, summer vibrancy, cinematic".

Figure 1: Two images generated by Stable Diffusion [RBL+22] .

Recently, advances in text-to-image synthesis have enabled people to create high-fidelity vivid-looking images by simply prompting the model with text descriptions [RDN+22]. However, learning to write the right prompt for creating an ideal image can hard. For example, Figure 1b is an carefully-engineered prompt with its corresponding output by Stable Diffusion [RBL+22], while as illustrate in Figure 1a, a naive prompt will usually lead to degradation in image quality. In fact, writing good prompts itself has been viewed as a professional skill as prompt engineering [LYF+22] and high-quality prompts have started to be sold on market places [1]. In this project, we aim to guide the user on the process of prompt creation, ease the human engineering efforts in creating the good prompt for Stable Diffusion by

1. Collecting a diverse set of highly rated prompts from Stable Diffusion community, which will serve as the dataset of high quality prompts for our research.
2. Creating a model that given a prompt as input, outputs modification suggestions (several modified prompts) that potentially make the prompt better. A prompt is evaluated by human examining the image generated by Stable Diffusion using this prompt. A good image should align with the input text and have high quality.

[1] https://promptbase.com/

2 Related Works

2.1 Text-to-Image Synthesis

Over the past few years, generative models are under fast development in the areas of graphic design, music production, article writing. The field of text-to-image synthesis using deep learning has recently become a hot topic due to diffusion models [DN21]. Diffusion models greatly improves the quality of image generated by GAN, which dominated generation tasks at that time. Now, more and more models, such as Stable Diffusion [RBL+22] and DALL-E [RDN+22], are released to public and have their own communities for designers or artists sharing their generated image. Figure 1 shows a sample generation result of Stable Diffusion.

2.2 Prompt Engineering

Prompt is the natural language used to specify the task users want to complete and is given to a pre-trained language model to interpret and complete. Prompt engineering is the process of creating a prompt that results in the most effective performance on the downstream task [LYF+22]. With zero shot capability of foundation models, prompt engineering has arisen across many areas, such as question answering [KMK+20], multi-modal learning [TMC+21], text-to-image synthesis.

In some areas, the prompt is designed by professionals and fixed for downstream tasks but prompts in text-to-image synthesis will be created purely by end users. Generated image quality can be varied vastly according to the prompt. We can see that figure 1b looks better than 1a by extending the prompt. Therefore, assisting end users in the prompt creation process is an emerging new research field. Recent works include guidelines [LC21] that help users to write better prompt for text-to-image generative models. Different types of prompt modifiers [Opp22] are classified and verified their utility by comparing the similarity between the output and the ground truth after iterative experimentation. The classification and the proof of necessity for constraints defines the problem more clearly, but it can not work out which prompt may lead to a better output, either. However, using a pre-trained language model to improve prompting itself in the area of text-to-image synthesis has huge potential to free users from prompt engineering and is largely unexplored. In this work, we aim to make one step to fill this blank.

3 Approach

3.1 Dataset Creation

Recent works have started mining high-quality prompts from websites. For example, the prompt dataset for the MagicPrompt Project[2] is crawled from Lexica.art. However, although websites like Lexica.art contains an extensive set of prompts. it has no guarantee on their quality. In contrast, we propose to mine high-quality prompts from the Stable Diffusion server on Discord, which contains rich user-feedback in the form of Emoji on each prompt. We will leverage these user feedbacks to rank and select high quality prompts. We expect the dataset to contain about ten thousand high-quality prompts.

3.2 Methodology

After obtaining the dataset, our task is to train a model that given a prompt as input, outputs modification suggestions that potentially make the prompt better. We model it as a sequence-to-sequence task and will finetune a pre-trained encoder-decoder based language models like BART [LLG+20] or T5 [RSR+20] for the job.

To model this task as a sequence-to-sequence problem, besides high-quality prompts as the output, we also need their corresponding short descriptive summary as input part of the data, which is used to simulate users' short input prompts. We plan to apply mature abstractive summary models like BRIO [LLRN22] to fulfill the task.

[2] https://huggingface.co/Gustavosta/MagicPrompt-Stable-Diffusion

3.3 Evaluation

Evaluation on the quality of open-domain synthesized image has always been hard. Recent works [RBL+22] have continued to use human evaluation for the analysis for 1.) the quality of image 2.) how well does the synthesized image align with the input prompt. We plan to follow prior works and conduct the evaluation experiment by:

1. Collecting a set of prompts from a broad audience (in reality, either friends or classmates).
2. Generating corresponding images, using Stable Diffusion, with text input both with or without modifications made from our model.
3. Conducting a double-blind experiment, evaluating which image among the two has better quality, and which has stronger alignment with the input text.

References

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[KMK+20] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November 2020. Association for Computational Linguistics.

[LC21] Vivian Liu and Lydia B. Chilton. Design guidelines for prompt engineering text-to-image generative models, 2021.

[LLG+20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[LLRN22] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. Brio: Bringing order to abstractive summarization, 2022.

[LYF+22] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, aug 2022.

[Opp22] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation, 2022.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[TMC+21] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models, 2021.