Prompt Assistant For Text-to-Image Synthesis Progress Report

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

1 Introduction

Text-to-image synthesis has been more and more popular recently thanks to the development of diffusion models (1). Users are able to use text to describe what they want, which is called prompt, and models like Stable Diffusion (6) or DALL-E (4) will return a synthesized image. However, thinking of a good prompt to generate an ideal image can be hard. Our project aims to help users write better prompts. In fact, how to write a good prompt has been investigated (3). One way is to provide a series of decorations and styles. Figure 1 shows such an example. We can see that although image 1a and 1b show the same content, image 1b has obviously higher quality by providing a series of decorations such as "portrait", "highly detailed" and so on. However, it is not obvious for a beginner to know which set of decorations fits their prompt best. In this project, we propose a model that predict a series of decorations to improve users' prompt.

(a) Image generated by plain text description: "Regal papal pond turtle wearing a pope hat".

(b) Image generated by enhanced text description with styles and decorations.

Figure 1: Two images generated by Stable Diffusion. Figure 1b is generated by prompt: "Regal papal pond turtle wearing a pope hat, d&d, fantasy, portrait, highly detailed, digital painting, trending on artstation, concept art, sharp focus, illustration, art by artgerm, Greg Rrutkowski and Magali Villeneuve pope francis, red ear slider".

2 Data Preprocessing

2.1 Dataset: DiffusionDB

DiffusionDB (10) is a large-scale text-to-image prompt dataset. It contains 14 million text-image pairs generated by Stable Diffusion using prompts and hyperparameters specified by real users. All

data are collected from the official Stable Diffusion Discord server. For our project purpose, we only need to use the prompt part the dataset.

2.2 Separating Decorations and Descriptions

Since our goal is to add decorations to description, we need to first divide our prompt dataset into descriptions and decorations. A prompt usually consist of multiple clauses (10) and descriptions and decorations are usually provided in different clauses. Since descriptions are usually longer, we classify clauzes with 6 words or more as descriptions and otherwise decorations. Take figure 1 as an example, the only description clause is "Regal papal pond turtle wearing a pope hat", which has 8 words. Obviously, this classification is not very accurate. "Greg Rrutkowski and Magali Villeneuve pope francis" also has more than 6 words but it is a decoration. Therefore, our description-decoration pairs have some noises. However, as our model can have decorations as an optional input together with descriptions, noise is not a very big problem. The problem is our model cannot generate decorations more than 5 words. We have intentions to improve this problem in the future.

Then we delete description-decoration pairs whose number of decorations is less than 6 and randomly select 0-3 decorations to concatenate descriptions to form an input and the rest of the decorations are output. This step is helpful for our model to find more decorations consistent with the decorations users provide. Thus, we get 600k input-output pairs to train our model and 100k pairs to evaluate. For example, an input could be "Regal papal pond turtle wearing a pope hat, Greg Rrutkowski and Magali Villeneuve pope francis, concept art, sharp focus" and the corresponding output could be "d&d, fantasy, portrait, highly detailed, digital painting, trending on artstation, illustration, art by artgerm, red ear slider".

3 Methodology

As both our input and output are string sequence, it is naturally to think of fine-tuning a pre-trained sequence-to-sequence model to solve our problem. One such model is BART (2).

3.1 BART

BART, developed by Facebook, is a transformer (8) model which uses 6 multi-head attention layers in the bidirectional encoder and 6 multi-head attention layers in the left-to-right autoregressive decoder. A larger BART version with 12 layers in each is also provided but that require more resources to fine-tune so we use the standard version. Encoder takes tokenized sentence as input and output an intermediate representation. Decoder takes the intermediate representation as input and auto-regressively predicts next token and feed the token back to decoder one by one.

BART is pretrained by first corrupting the text with arbitrary noise and then using the corrupted text to predict the original text. This training task enables BART to be trained on unlabeled dataset, which means the dataset can be extremely large. The model and pretraining is shown in figure 2.

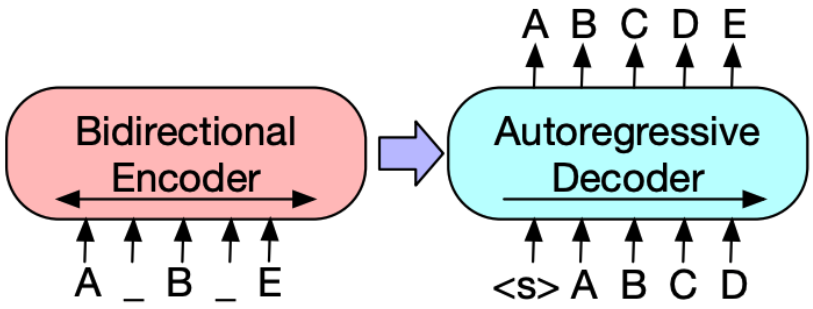


Figure 2: The Encoder takes an corrupted text as an input and output an intermediate representation to decoder. The decoder will receive a start signal <s> and output a token which will then feed back to decoder and repeat this step until decoder output an end signal.

3.2 Fine-Tuning

Though BART is trained on restoring corrupted text, it also learns text comprehension in this task so it can be particularly effective when fine tuned for other text generation tasks. Though BART itself is trained on an extremely large unlabeled dataset, we are able to use a relatively small labeled dataset, 600k in our case, to generate a fairly good model. In terms of training parameter, we choose batch size 8 and learning rate $5.6 * 10^{-5}$.

4 Current Result

It is hard to evaluate the quality of a prompt directly. Here, we assume a better image implies a better prompt. Therefore, we plan to use CLIP aesthetic score (9) and human evaluation to compare the images generated by original prompt and enhanced prompt. Those are going to be done in our next step and currently, we can show some examples.

Figure 3 shows the prompt and the image improved by our model compared to the original ones. We can see that our model makes the prompt emphasize on realistic and details, which fits well on the prompt content and thus Diffusion Stable is able to generate much better image. Figure 4 is another example. Our model provides more details to Stable Diffusion like "sunset" and "dramatic lighting" to generate a better image.

From above two examples, we can see that our model is able to find the best style for the given prompt and express the style as a set of decoration clauses. Although the image content is about the same, the image quality can be largely improved by our model's predicted decorations.

(a) Image generated by the original prompt.

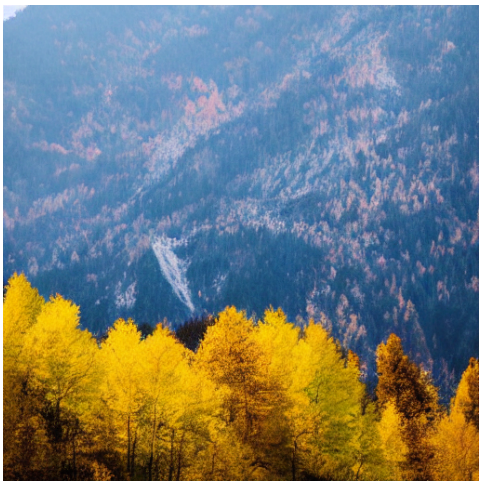
(b) Image generated by the enhanced prompt.

Figure 3: Two images generated by Stable Diffusion. Figure 3a is generated by the original prompt: "robot army rule the world". Figure 3b is generated by the prompt improved by our model: "robot army rule the world, ultra realistic, concept art, intricate details, highly detailed, photoreal".

5 Future Plan

5.1 Dataset Improvement

Current dataset label method labeled word groups with 5 or more words as description, otherwise labeled them as tag. However, the method sometimes fails because a few users use discrete words to describe the content of the image instead of using a complete sentence. To solve that, we propose to use the characteristic of words and decision tree to increase the accuracy of the method. At present, a subset of 400 word groups has already been labeled to be used to train our decision tree. Finishing training the decision tree, a comparison will be made between the original model and the new one.



(a) Image generated by the original prompt.

(b) Image generated by the enhanced prompt.

Figure 4: Two images generated by Stable Diffusion. Figure 4a is generated by the original prompt: "mountains with yellow leaves". Figure 4b is generated by the prompt improved by our model: "mountains with yellow leaves, mountains in background, sunset, dramatic lighting, artstation, mat".

5.2 Different pre-trained models

Besides BART-base, we also plan to re-train our model in T5-base (5), which is an encoder-decoder Language model pre-trained on multi-task. The different pretraining objective might lead to different performance in our task which we plan to evaluate.

5.3 Human evaluation

We plan to conduct an experiment to evaluate our model, which means we need to whether original prompt or the prompt improved by our model can generate a better image. Recent works (6) 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.

5.4 CLIP Aesthetic Score

Besides costly human evaluation, we also plan to use CLIP Aesthetic Score (7), a machine-based aesthetic evaluation metric. We plan to follow the procedure detailed in Section 5.3 but instead, use CLIP Aesthetic Score to evaluate which image has better quality.

References

[1] DHARIWAL, P., AND NICHOL, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems* (2021), M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., pp. 8780–8794.

[2] LEWIS, M., LIU, Y., GOYAL, N., GHAZVININEJAD, M., MOHAMED, A., LEVY, O., STOYANOV, V., AND ZETTLEMOYER, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL* (2020).

[3] LIU, V., AND CHILTON, L. B. Design guidelines for prompt engineering text-to-image generative models, 2021.

[4] RAMESH, A., DHARIWAL, P., NICHOL, A., CHU, C., AND CHEN, M. Hierarchical text-conditional image generation with clip latents, 2022.

[5] ROBERTS, A., AND RAFFEL, C. Exploring transfer learning with t5: the text-to-text transfer transformer. *Accessed on* (2020), 23–07.

[6] ROMBACH, R., BLATTMANN, A., LORENZ, D., ESSER, P., AND OMMER, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 10684–10695.

[7] SCHUHMANN, C., BEAUMONT, R., VENCU, R., GORDON, C., WIGHTMAN, R., CHERTI, M., COOMBES, T., KATTA, A., MULLIS, C., WORTSMAN, M., ET AL. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

[8] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U., AND POLOSUKHIN, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[9] WANG, J., CHAN, K. C. K., AND LOY, C. C. Exploring clip for assessing the look and feel of images, 2022.

[10] WANG, Z. J., MONTOYA, E., MUNECHIKA, D., YANG, H., HOOVER, B., AND CHAU, D. H. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]* (2022).