Improvement on Prompt-to-Prompt Image Editing

Yanchong Peng, Yuhe Peng, Peng Huang

{ycpeng, pengyh, phuang}@bu.edu

Abstract

This report covers two improvements we made to the prompt-to-prompt image editing algorithm by the milestone. The original technique has the limitation of generating both the source and edited images with one random seed, which does not work out when we mix the features of two images generated by two random seeds. So we add one more seed to support this. Another improvement is that now we can remove a specific target from the source image by scaling the attention map of the object.

Introduction

Many techniques have been proposed in text-to-image, such as Imagen(Saharia et al. 2022) and DALL-E2(Ramesh et al. 2022). However, these models cannot control the global sections of the image; that is, if changing a tiny part of the text prompt to another, such as ”dog” to ”cat”, the whole generated image may not be the same, not only ”dog” but also the background.

Related Work Bau(Bau et al. 2021) showed that the mask provided by the user could help edit the text-based image locally. This mask contains all the areas relevant to the text that the user wants to change, whereas it requires lots of labor and time, which is inefficient. Seminal works combining GAN(Goodfellow et al. 2020; Brock, Donahue, and Simonyan 2018), primarily used to generate high-resolution images, and CLIP(Radford et al. 2021), which contains rich image-text representation, do not require extra labor annotations. Nevertheless, they cannot handle vast and varying datasets.

Text2LIVE(Bar-Tal et al. 2022) also can edit text-based localized images. However, this model needs to apply the network to each input, which is highly time-consuming, and can only change the textures, not the whole patterns, such as a plane to a train.

Hertz(Hertz et al. 2022), whose paper this project is relevant to, modifies the internal attention maps of the diffusion model, which not only preserves the unchanged sections of the origin but also modifies the part that the user wants to replace. This way, we just need the text input and do not require extra labor. The Hertz team has released the newest version of the prompt-to-prompt model using pre-trained huggingface/diffusers models. They implemented the method based on Latent Diffusion and Stable Diffusion, respectively. We will produce our work based on the most updated GitHub repository.

Goal For the current stage, this technique is subject to several limitations. First, there is only one random seed s for both source and edited image, which means it is impossible to keep the images for both source and edited that randomly appear. Second, although it provides a way to reduce global noise(-rocks, -fog), it does not have the functionality of removing a specific target. Moreover, the method cannot change an object's location in the picture. Finally, the algorithm for sequential image editing is inefficient.

So, in this project, we expect to make some improvements to the existing algorithm to achieve the following goals:

- Add one more random seed s^* to generate and keep source/edited images separately.
- Remove a specific target in the image.
- Combine two specific images. (Move the car generated by random seed s into the street generated by seed s^*).
- Move a specific object across the image.
- Improve the algorithm performance in continuous editing.

By the end of today, we have finished the first two sub-goals. Detailed metrics will be provided in the later sections.

Methods

Formulation

The source image $\mathcal{I}$ is generated by a text-guided diffusion model with prompt $\mathcal{P}$ and random seed s. The goal is to edit the source input image with edited prompt $\mathcal{P}^*$, and another random seed s^*.

Model and Dataset

We use the pre-trained model Stable Diffusion v1-4(Rombach et al. 2022) as the backbone of this algorithm.

This model was trained with 256*256 resolution on the laion2B-en dataset and 512*512 resolution on the laion-high-resolution dataset, which contains over 170M images with 1024*1024 resolution.

Algorithm 1: Multi-seed image editing

```
**Input**: A source prompt $\mathcal{P}$ with a random seed $s$, a target prompt $\mathcal{P}^*$ with a random seed $s^*$
**Output**: A source image $x_{src}$ and an edited image $x_{dst}$
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed $s$
$z_T^* \sim N(0, I)$ a unit Gaussian random variable with random seed $s^*$
for $t = T, T-1, ..., 1$ do
    $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t)$
    $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t)$
    $\widehat{M_t} \leftarrow Edit(M_t, M_t^*, t)$
    $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t)\{M \leftarrow \widehat{M_t}\}$
end for
return $(z_0, z_0^*)$
```

This model used the ViT-L/14 text-encoder to encode text prompts and encoded image prompts into latent representations. Later, with cross attention, the model injected the output of the text-encoder into the UNet(Ronneberger, Fischer, and Brox 2015) in the diffusion model.

Methods

- **1. Multi-seed Image Editing (accomplished)** The input of our modified algorithm has one more random seed than the original algorithm: one source random seed s and one target random seed s^*, as described in Algorithm 1. Furthermore, since the initial latent image can entirely determine the randomness, there is no need for the diffusion model to accept any random seed anymore, which simplifies the design.(Hertz et al. 2022)
- **2. Removing Specific Target (accomplished)** The user can remove the object from the image generated before. For instance, the source text prompt is ”a car parking on the street”. If the user wants to remove the car, let the attention map of the token ”car” as j^*. Then, the maps-editing operation function will be modified as

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} -5 \cdot (M_t)_{i,j} & \text{if } j\text{=}j^* \\ (M_t)_{i,j} & \text{otherwise} \end{cases}$$

  The weight -5 is the result after several experiments. With this scale, the output image performs more natural and adaptive.
- **3. Combining Two Images** We proposed an extended application of combining two source images based on the original algorithm (Hertz et al. 2022) that only takes one source image. Suppose we have source prompts $\mathcal{P}$ and $\mathcal{P}'$, and a target prompt $\mathcal{P}^*$, a combination of $\mathcal{P}$ and $\mathcal{P}'$. In a single step t of the diffusion process, we get the attention maps M_t, M_t', M_t^* respectively for prompts $\mathcal{P}$, $\mathcal{P}'$, $\mathcal{P}^*$, and put them into the edit function

$$(Edit(M_t, M_t', M_t^*, t))_{j^*} = \begin{cases} (M_t)_{j^*} & \text{if } j^*\text{=}j \\ (M_t')_{j^*} & \text{if } j^*\text{=}j' \\ (M_t^*)_{j^*} & \text{otherwise} \end{cases}$$

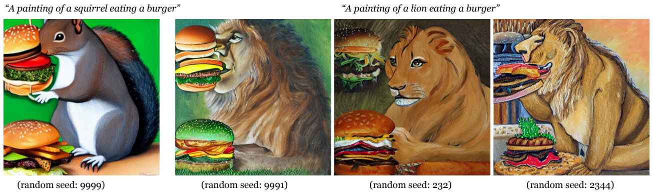
"A painting of a squirrel eating a burger"
"A painting of a lion eating a burger"
(random seed: 9999)
(random seed: 9991)
(random seed: 232)
(random seed: 2344)

Figure 1: The results of multi-seed image editing. We change the squirrel to a lion with the same target prompts but different random seeds. Attention control is employed to try to make other parts unchanged. The overall composition of these edited images almost remains the same.

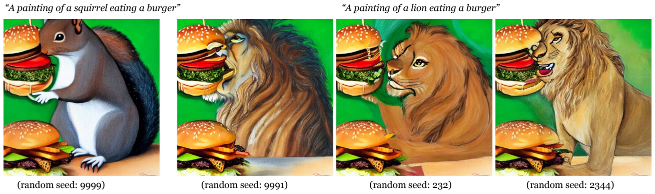
"A painting of a squirrel eating a burger"
"A painting of a lion eating a burger"
(random seed: 9999)
(random seed: 9991)
(random seed: 232)
(random seed: 2344)

Figure 2: The results of multi-seed image editing with mask. We change the squirrel to a lion, with the same target prompts but different random seeds. Attention control as well as latent image masking is employed to try to make other parts unchanged. The other elements of these edited images almost remain identical.

  where j, j', and j^* are the text tokens for M_t, M_t', and M_t^* here. In fact, we can extend the combination from two images to multiple images by using a similar way.
- **4. Moving Object** We proposed a method of moving an object across an image, which initially serves as a limitation in the paper by (Hertz et al. 2022). For precise moving, we accept a target location R as the destination of moving and a prompt token for the object we would like to move. We define a move function $Move(M_t, R)$ to move the focus area of M_t to the target location R. Hence, the edit function should be defined by

$$(Edit(M_t, t))_j = \begin{cases} Move((M_t)_j, R) & \text{if } j \text{ is to move} \\ (M_t)_j & \text{otherwise} \end{cases}$$

- **5. Continuous Editing** We will improve the algorithm (Hertz et al. 2022) for efficient continuous editing performance by saving all the latent edited attention maps $\widehat{M_t}$ for prompt $\mathcal{P}^*$. When the target image z_0^* serves as a source image of another continuous editing, we input $\mathcal{P}^*$ and pre-saved $\widehat{M_t}$ into the algorithm to achieve efficient computation without repeat. We can also leverage this method in other scenarios, like combing two images to reduce redundant computation.

Results

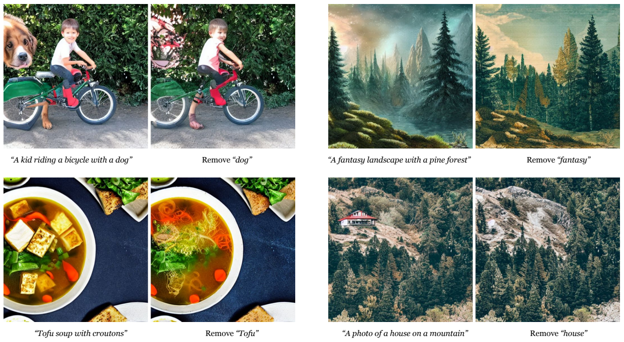
"A kid riding a bicycle with a dog"
Remove "dog"
"A fantasy landscape with a pine forest"
Remove "fantasy"
"Tofu soup with croutons"
Remove "Tofu"
"A photo of a house on a mountain"
Remove "house"

Figure 3: The results of removing object experiments. The top right one is considered from global modification, and others are traditional target replacements. Except for the top left one, our remove method does a great job on multiple text prompts. The algorithm does not modify irrelevant objects in their attention maps.

Multi-seed Image Editing We tested multi-seed image editing with the application of target replacement. The results are displayed in Figure 1. We replaced squirrel with lion, and used the same target prompts with three random seeds. We can observe that the algorithm returned lions in different styles while maintaining the source image's overall composition and structure, which indicates our improvement's initial success.

However, we can also notice that the styles of burgers and backgrounds also changed. In order to fix this problem, we ever tried to apply the mask to the initialization of the initial latent image to retain the unchanging area. However, in that method, the outcome was unrecognizable. Therefore, we resorted to applying the mask to intermediate latent images instead of the initial noisy image. As displayed in Figure 2, the results of applying the mask show that the styles of background and burgers almost remained the same, demonstrating that our conception is practical.

Remove Object We perform this feature in two different styles: target replacement and global modification. The results are displayed in Figure 3. The top right one is considered from global modification, and others are traditional target replacements.

Except for the top left one, our remove method does a great job on multiple text prompts. The algorithm does not modify irrelevant objects in their attention maps.

Github Repository

https://github.com/ycpeng8/improvement_p2p

References

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2LIVE: Text-Driven Layered Image and Video Editing. *arXiv preprint arXiv:2204.02491*.

Bau, D.; Andonian, A.; Cui, A.; Park, Y.; Jahanian, A.; Oliva, A.; and Torralba, A. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.