Dreambooth in Action: Personalized Text-to-Image Generation

DDA4210 Advanced Machine Learning Course Project
Instructor: Jicong Fan

Zongbo Bao🐾, Yu Lu🐾, Lian Zhong🐾, and Jiahui Xu🐾

🐾*The Chinese University of Hong Kong, Shenzhen*

I. Introduction

Recent advancements in artificial intelligence have seen significant developments in text-to-image generation models. Key players like OpenAI's DALL-E 2, Google Brain's Imagen, and Stability AI's Stable Diffusion have marked a new era in generating photorealistic images from textual descriptions. While these text-to-image models are capable of generating high-quality images, they still face challenges in accurately reproducing specific subjects from provided images and placing them in new, diverse contexts in the generated images.

Addressing this gap, Dreambooth [1] presents a unique solution. It goes beyond generic image synthesis, equiped with the ability to retain the distinctive features of a particular subject. By fine-tuning the base model with only several training images, Dreambooth learns to associate a unique identifier with a specific subject. This process enhances the model's capacity to generate novel images of the subject, maintaining their key characteristics across different scenes and settings.

In our work, we first implement the Dreambooth on a dataset of only 4 pictures in a selected subject such as dog, cat or any other entity. Then, a *unique identifier* is used to notify the class-specific instance, effectively teaching the model to recognize and replicate the specific characteristics of the subject. Lastly, a *prior preservation loss* is added in the model to solve the problem of *language drift* and *low diversity*. We apply the models trained on different subjects to several tasks, including *subject recontextualization*, *property modification* and *artistic rendering*, ensuring the preservation of proper portion of subject's key features.

II. Methodology

The basic idea of the model training is to input few pictures in a specific subject, aiming to generate new images with high quality.We proceed by outlining our approach, including an overview of pre-trained diffusion models, the fine-tuning process for subject identification, and the implementation of a class-specific prior-preservation loss to mitigate language drift and low diversity problems.

A. Pre-trained Diffussion Model

Stable-diffusion-2 (model repository) is selected as our pre-trained model. The architecture of pre-trained model can be found in [2].During fine-tuning, the pre-trained model $\hat{\mathbf{x}}_\theta$, is given an initial noise map $z_t := \alpha_t x + \sigma_t \varepsilon$. A text encoder Γ and a text prompt P are adopted to generate a conditioning vector $c = \Gamma(P)$, which encodes textual information of prompt and guides the de-noising steps. After the image $x_{\text{gen}} = \hat{x}_\theta(\varepsilon, c)$ is generated, pixel-wise squared error is used. Normal fine-tuning loss of the model is listed below:

$$\mathbb{E}_{\mathbf{x},c,\varepsilon,t}\left[w_t \left\|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \varepsilon, c) - \mathbf{x}\right\|_2^2\right] \tag{1}$$

B. Personalization of the Diffusion Model

Setup for Prompts The goal of our project is to input a (unique identifier, new object) pair to let the model "learn" the key features of the new object. We used a simple approach to label all the image with the same text: "a [identifier][class noun]" as shown in the upper half part of Figure 1. Here identifier is used to reference the new concept and the class noun is a general class description of the object (e.g. dog, cat, toy, etc.).

Unique Identifier Choosing the suitable unique identifier is essential for training. Common English word like "unique" or "special" is not recommended because the model has already had prior knowledge of the words. It becomes difficult for the model to use these words to reference the new instance. We followed previous works and used the word "sks" as the unique identifier. Later, we show that a short sequence of character (3-5 words) that the model does not have prior semantic knowledge of is suitable as the unique identifier.

C. The incorporation of Prior Preservation Loss

Two issues will emerge as a result of fine-tuning all layers of the basic model. One is Language Drift, which is the semantic knowledge of the subject is lost during the process of training. For example, when taking 4 pictures of 'a sks dog' as training pictures, all the prompt of generating a normal dog will tend generate 'sks' dog, which means the meaning of a 'dog' is lost. The other issue is the low diversity in the output poses and views of the subject. To solve the two problems, a prior preservation loss is introduced in the

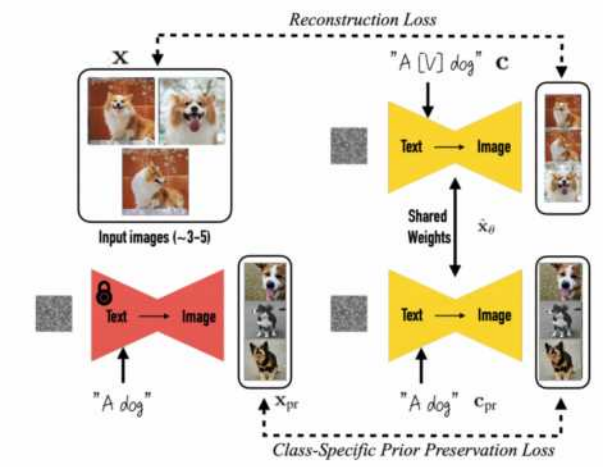

Fig. 1: **Fine Tuning** We fine tune the model using new object's images with prompt "a [identifier] dog". At the same time, we user a prior preservation loss to encourage the model to generate diverse images of the class (dog) images by inputting various dog images with prompt "a dog".

pre-trained model. By applying the ancestral sampler to the static pre-trained model to create data$x_{\text{pr}} = \hat{x}(z_{t1}, c_{\text{pr}})$, using random initial noise $z_{t1} \sim \mathcal{N}(0, I)$ and a conditioning vector $c_{\text{pr}} := \Gamma(\text{"a [class noun]"})$. The loss function is then listed as follows:

$$\mathbb{E}_{x,c,\varepsilon,\varepsilon',t}\Big[w_t \left\|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \varepsilon, c) - \mathbf{x}\right\|_2^2 + \lambda w'_t \left\|\hat{\mathbf{x}}_\theta(\alpha'_t \mathbf{x}_{\text{pr}} + \sigma_t \varepsilon', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\right\|_2^2\Big] \tag{2}$$

The basic idea keeps the same except the prior-preservation term in after the previous model structure. This loss supervises the model with its own generated images, and manages for the relative weight of this term.

III. Results

Three models are fine-tuned based on the Stable Diffusion model, with 2 instances: a corgi and a cat from the upper campus of the Chinese University of Hong Kong, Shenzhen.

A. The model of 'sks dog' without prior preservation

The first model is trained without the prior preservation, using 4 images of a corgi. The corresponding prompt is 'a sks dog'. Figure 2 shows the training set.

Figure 3 shows some results of the first model.

From the results, low diversity can be observed, as the dog always has the same action, and the images are always frontal view, even if we use prompts with different actions. Furthermore, there is an obstacle of language drift, which means when fine-tuning a model, it loses the prior semantic information it has learned. When prompted with 'a dog', the fine-tuned model only generates images of the specific dog used during training, rather than representing various breeds of dogs from different viewpoints and backgrounds.

Fig. 2: The training set of 'a sks dog'.

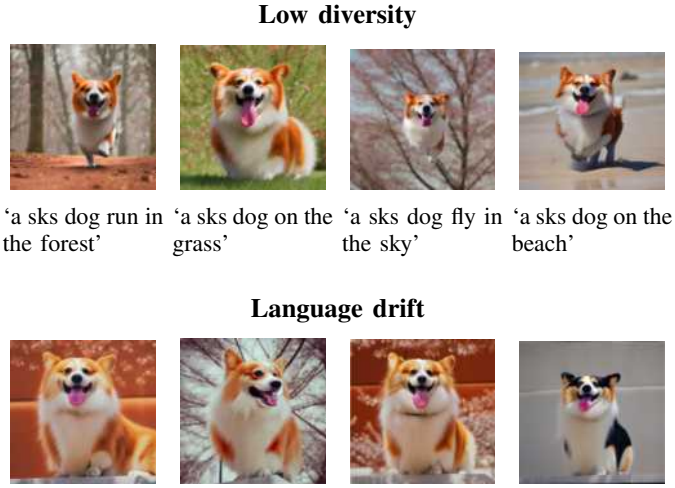

Fig. 3: Results of the model without prior preservation.

B. The model of 'sks dog' with prior preservation

The second model is trained with the prior preservation. Besides using previous four images of the corgi, the pre-trained Stable is applied to generate 12 images of 'a dog', which are also included in the training set and serve the purpose of supervising the fine-tuned model to preserve the concept of 'a dog'. Figure 4 shows some examples of images generated by the pre-trained Stable Diffusion and results of the second model.

As displayed in Figure 5, the model successfully generates novel and distinct image of dogs, which indicated the problem of language drift is solved. Furthermore, there is an increased diversity in contexts and viewpoints. The 'sks dog' in output images exhibits various actions, backgrounds and viewing

Fig. 4: Images generated by the pre-trained Stable Diffusion.

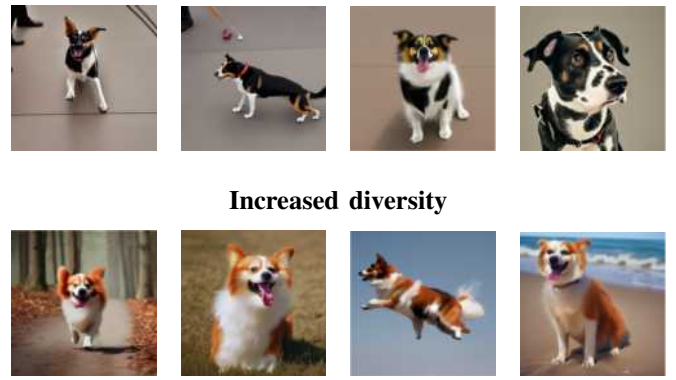

Fig. 5: Results of the model without prior preservation.

angles.

However, the second model still exists issues. One is that the generated dog always has its mouth opened, which is probably attributable to the limited diversity of the training set. Examples are in Figure 6.

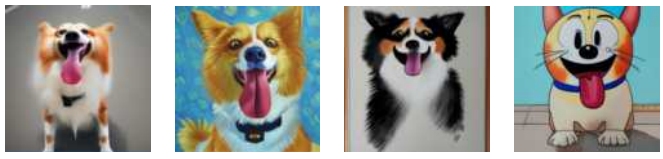

Fig. 6: A 'sks' dog always with its mouth opened.

Fig. 7: Problem with the unique identifier 'sks'.

Fig. 8: The training set of 'a mscds cat'. The cat's name is Chao Da Sheng and lives at Muse College in upper campus. A very lovely cat :).

Another one is the choice of the unique identifier. As shown in Figure 7, when utilizing the model to generate the cross of 'a sks dog' and a panda, the output contains a gun. The reason is that the pre-trained model has the prior semantic knowledge of the word 'sks', which refers to a gun.

C. Second Instance and Applications

As an experiment, the model is trained on the second instance which is a cat from CUHKSZ's upper campus (Figure 8). The training dataset contains 4 photos of the cat with various angles and distances. The unique identifier is changed to "mscds" which the model does not contain prior semantic knowledge for. We observed that the key feature, which is the white part above the cat's nose, is preserved in most of the generated photos and not issues due to prior knowledge in the unique identifier. In this section, we show some of the applications of our fine-tuned model.

Fig. 9: **Recontextualization**

Chinese Starry Night Mona Lisa Claude Monet

Fig. 10: **Art Rendition**

- **Recontextualization** With descriptive prompts, our model is able to generate pictures of the instance in new contexts (Figure 9). Specifically, the model is able to generate images of new poses of the instance such as "making a cake" or "playing the piano", and new facial expressions unseen from the training photo such as opening mouth and closing eyes. We also observe interaction of our instance and the scene like shadow and contact between the instance and objects.
- **Art Rendition** the model is able to generate images of our instance is a specific artistic style if we give the prompt " a painting/ drawing of a mscds cat in the style of [a famous artwork]" (Figure 10). Unlike style transfer, where only the style of the picture changes but the overall structure remains unchanged. Our model is able to generate various images of the instance.
- **Accessories** with the prompt "a mscds cat wearing a [outfit description] outfit,", we are able to generate photos of our instance in different outfits(Figure 11). We observe realistic interactions between the instance cat and the accessories.

CONCLUSION

In our project, we trained the Dreambooth model on the subjects of corgi dog and muse cat respectively. For each of these subjects, we utilized a minimal training set comprising only four images, which is a highly accessible model for people to get their desired images with the unique characteristics from the original dataset. The incorporation of prior preservation loss was pivotal, ensuring that the generated images maintained proper proportions and exhibited a rich variety, avoiding the drawback of low diversity often seen in similar models. Our Corgi dog model and muse cat model are available on Huggingface. Please visit dog model link or cat model link for further details.

Christmas Police Chef Halloween

Fig. 11: **Accessories**

REFERENCES

[1] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[2] 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*, pages 10684–10695, 2022.