Comics Illustration Synthesizer using Deep Generative Models

Mahmoud ElAref
Supervised by Prof. Lydia Chen, Dr. Zilong Zhao
Delft University of Technology

Background

- General Diffusion Models: are machine learning systems that are trained to denoise random gaussian noise step by step, to get to a sample of interest, such as an image [1].
- Stable Diffusion is is a latent text-to-image diffusion model. Meaning it does this in latent space.

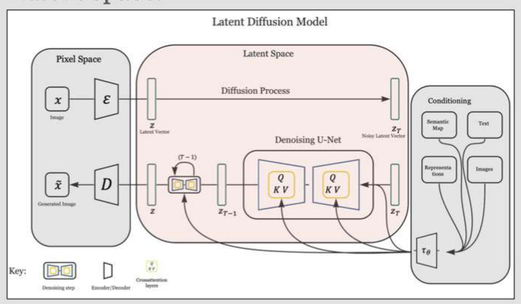

Figure 1: Latent Diffusion models

- Dreambooth is a method of fine-tuning text-to-image models that learns how to bind a unique identifier with a specific subject. [2]
- LoRA: (Low-Rank Adaptation of Large Language Models) is a novel technique introduced by Microsoft researchers to deal with the problem of fine-tuning large-language models. [3]
- LoRA can be applied to the cross-attention layers in text-to-image models.

Research Question

- "How can the Stable Diffusion model be fine-tuned such that it generates high quality Dilbert comic illustrations from text descriptions?"
- "How do the results of the fine-tuned Stable Diffusion model compare to those of the conditional GAN model in terms of quality and accurately matching preconditions"
- "Which fine-tuning method, Dreambooth or Lora, produces the highest quality Dilbert comic panels and accurately conveys the textual descriptions?"

Motivation

- Comics synthesis has struggled so far
- Comic synthesis can be potentially improved if latent diffusion models were used instead of conditional generative adversarial networks.
- Little resources comparing the fine-tuning techniques, Lora and Dreambooth

Methodology

I. Dreambooth vs Lora

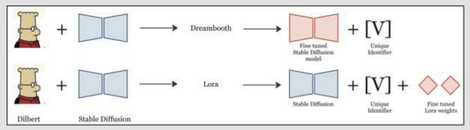

Figure 2: Diagram illustrating the difference in the model output of Dreambooth and Lora

- Dreambooth modifies the whole stable diffusion model and outputs a new one
- Lora instead of updating the whole model, it adds new weights inside it, without changing existing weights in the model

II. Experiment:

- Training the same base model using both techniques.
- Training for the same time, in the same environment and using the same dataset.
- Comics produced with both models, inferred under controlled conditions will be compared.

III. Dataset

- Dataset of 180 image-text pairs
- Cropped to separate panels and dialogue was removed.
- Manual captions outperformed automated BLIP captions.
- Captions included character names and all present objects or actions separated by commas.

Figure 3: Sample from the dataset, captioned: "Dilbert, Boss, Boss sitting at table, Dilbert sitting at table, Boss holding paper, Dilbert holding phone, coffee cup, table"

Results

I. Dreambooth

Figure 4: Results from Dreambooth fine-tuning produced FID score of 154

II. Lora

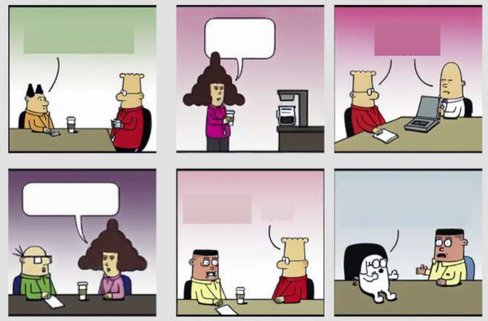

Figure 5: Results from Lora's fine-tuning produced FID score of 123

III. Comparison

Model	Chr. 1	Chr. 2	Perfect scene
Dreambooth	98%	18%	15%
Lora	91%	77%	53%

Table 1: Results for each model in the accuracy experiment. Perfect scene: Both characters present alone without any extra characters

Contact

Name: Mahmoud Elaref
Email: m.m.elaref@student.tudelft.nl
Responsible Professor: Prof. Dr. Lydia Chen
Supervisor: Dr. Zilong Zhao

Limitations and Future Work

- While Lora was able to adapt to the Dilbert style, Dreambooth lost touch with the style in some scenarios like outside the office or without the main character Dilbert.
- Dreambooth is used to smaller datasets, Lora was able to cope with the large number of characters while Dreambooth struggled in differentiating certain characters.
- Future work: Noe image models can produce comic panels without dialogue. How can LLM be incorporated to generate text that can be added on top of the images to fully automate the comic generation process?

Conclusion

- Lora outperformed Dreambooth in terms of quality, it managed to follow the text prompts more accurately, producing perfect scenes more than once every two panels.
- Lora outperformed Dreambooth in terms of image quality and similarity to the original Dilbert comics, with an FID score of 122.89
- Stable Diffusion produced Dilbert comics that were of significantly higher quality and greater detail than conditional GANs

References

[1] Robin Rombach et al. High-resolution image synthesis with latent diffusion models, 2022.
[2] Nataniel Ruiz et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.
[3] Edward J. Hu et al. Lora: Low-rank adaptation of large language models, 2021.