Machine Learning Homework 3

Gausse Mael DONGMO KENFACK (董沫高斯)
Student ID: 2024403346

January 2025

Contents

Introduction

The generative AI in today's fast-evolving technological landscape stands at the very forefront of innovation. Understanding how and on what principles generative AI works is rewarding and very instrumental to AI researchers. The applications of generative AI vary from text-to-text generation, further to text-to-image synthesis, and even the creation of videos from text. Among them, text-to-image generation attracts considerable attention from both research and industry since it can generate amazing, highly detailed visuals from natural language textual prompts. One popular open-source tool in this domain is Stable Diffusion [3] [4], a state-of-the-art text-to-image generation model. Stable Diffusion is based on the concept of diffusion models, which generate images by iteratively adding noise and then denoising this noisy input through the careful design of a noise schedule and sampling algorithm.

Not everything that generative AI does needs to involve creating something from scratch. More often, a particular artistic style or aesthetic will be needed, developed from an already existing visual inspiration. This is formally known as style transfer, a process usually consisting of fine-tuning the underlying generative model toward adapting to the style desired, keeping the same level of understanding and response to the prompts of the model. The aim of this homework is to study in detail how to fine-tune Stable Diffusion for style transfer.

1 Fine-Tuning Process

1.1 Stable Diffusion

Stable Diffusion is a state-of-the-art generative AI model for text-to-image synthesis, based on the concepts of latent diffusion models. It refines noise into coherent and detailed images in progressive steps, conditioned on a textual prompt. The structure will be composed of several key components that fit together. At its core is the U-Net architecture, a neural network responsible for denoising the latent representations at each step of the diffusion process. The U-Net is conditioned on text embeddings generated by the CLIP text encoder, which translates natural language prompts into a form the model can interpret. Other important ingredients are VAEs, which map images from pixel space into a compressed latent space for computational efficiency and effective reconstruction. The scheduler or noise scheduler controls the process of diffusion by how much noise to add or remove at what step, keeping the balance for stability, hence control in the generation process. Also, Stable Diffusion supports fine-tuning using methods such as LoRA [1], enabling it to learn a given style or an aesthetic relevant to a domain with minimal computational cost. These together empower Stable Diffusion to make high-quality, contextually relevant, and visually appealing images; thus, it becomes quite versatile in the domain of generative AI.

1.2 LoRA

LoRA stands for Low-Rank Adaptation and is one of the fine-tuning approaches that are much more efficient. LoRA focuses on adapting large pretrained models for new tasks or styles, where a full retraining would be cumbersome or not required. Instead of updating all model parameters, LoRA injects low-dimensional, task-specific parameter matrices into the architecture that get updated, whereas the weights of the original model remain untouched. These extra matrices approximate the update of the parameters in a low-rank subspace with significantly less computation and memory overhead in comparison with full parameters fine-tuning. This property especially makes LoRA well adapted to resource-sensitive applications, where a single base model may have to be fine-tuned onto multiple specialized tasks. This gives the Stable Diffusion model an opportunity to study new artistic styles or domain-specific aesthetic features without interfering with its general skills of text-to-image generation. After being fine-tuned, these LoRA parameters could then be dynamically used to manipulate the model with flexibility and reusability without touching the integrity of the source model. With LoRA, one can achieve state-of-the-art style transfer or task adaptation in a very efficient and scalable manner.

1.3 Implementation

There exist a lot of LoRA fine-tuning implementations online [5] but for this homework the goal was to implement and debug the training loop manually and make it work.

We ran the code for 1000 steps with same hyper-parameters on 2 different datasets:

- The first one is the ***Doodle dataset*** [2], we found it on ***huggingface***, it's a collection of 1000 doodles with text captions, you can see a sample in Figure 1.
- The second one that we also found on ***huggingface*** is the ***Vintage dataset*** [6], it has 685 images and a sample is on Figure 2

Figure 1: A Sample of the doodles dataset

Figure 2: A Sample of the vintage dataset

2 Analysis of the fine-tuned Model

2.1 Generating some styled Images

After successfully fine-tuning the model, we generated some images and we can see that the generated images have the style of the ones from the datasets.

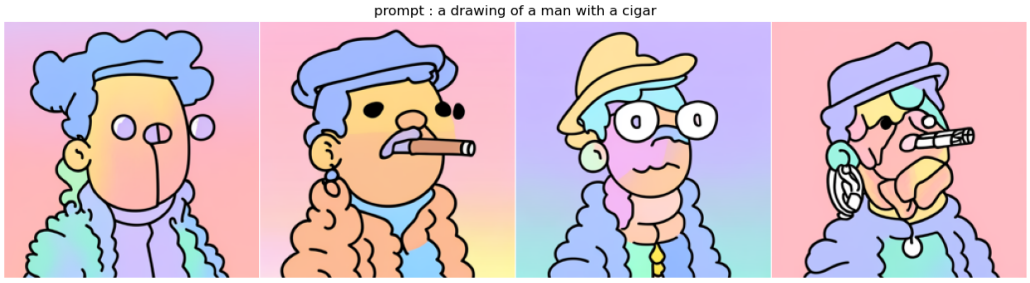


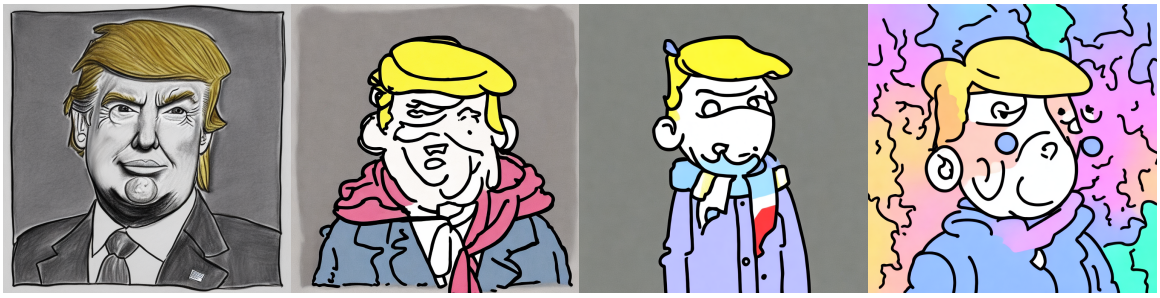
Figure 3: Some generation results from doodles

Figure 4: Some generation results from vintage

Figure 5: Some generation results from doodles

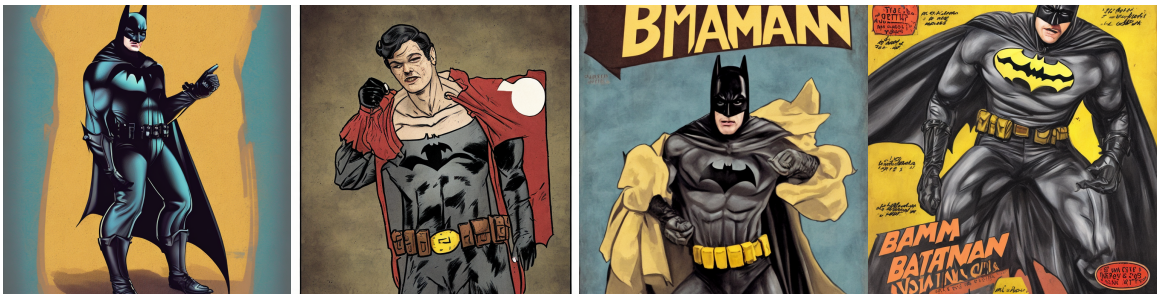
2.2 Evolution of generations During the training

While running the training process, we tracked it with ***Weights and Biases*** [7] and every 2 epochs we generated some images as during the validation process. in this section we will show (6, 7) and comment the evolutions of the fine-tuned model. We can see that the models progressively capture the particularity of the images, like the colors for "doodles" or the text present on images for "vintage".



(a) 84/1000 steps (b) 252/1000 steps (c) 420/1000 steps (d) 588/1000 steps

Figure 6: Evolution of the vintage fine-tuning with the prompt : a drawing of Donald Trump with a scarf



(a) 58/1000 steps (b) 290/1000 steps (c) 522/1000 steps (d) 870/1000 steps

Figure 7: Evolution of the vintage fine-tuning with the prompt : Batman as a sherif in a vintage style

2.3 Fine-Tuned vs original Stable Diffusion

In this section we generated some images with the fine-tuned models and with the original stable diffusion with same prompts, let's take a look.

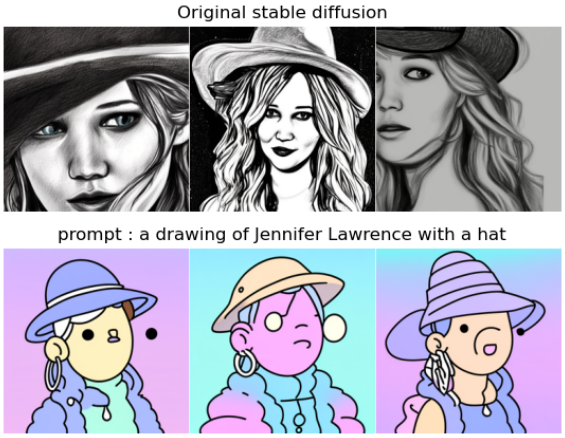


Figure 8: Comparison with doodles

Figure 9: Comparison with vintage

Conclusion

In conclusion, the fine-tuning process has demonstrated remarkable effectiveness, as evidenced by the high-quality results obtained. The incorporation LoRA played a pivotal role in streamlining the process, making it not only efficient but also computationally lightweight. The ability to achieve such impressive outcomes within a relatively short training time highlights the power of LoRA in adapting large models like Stable Diffusion to specialized tasks.

Looking ahead, there are several promising avenues for further exploration. One important aspect to investigate is the potential for overfitting during the fine-tuning process. Additionally, experimenting with alternative hyperparameters—such as learning rates, batch sizes, LoRA rank, gamma SNR may reveal configurations that yield even better results or improve training stability.

Another area worth exploring is the use of diverse datasets for fine-tuning. By applying the model to datasets with varying styles, content, or domains, it would be possible to evaluate its adaptability and robustness across different tasks. This could also uncover limitations or biases in the current fine-tuning approach, paving the way for further enhancements.

References

[1] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685 (2021). arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.

[2] Julian Moras. *Doodles dataset*. https://huggingface.co/datasets/julianmoraes/doodles-captions-BLIP. [Online].

[3] *Stable Diffusion 2.1*. https://huggingface.co/stabilityai/stable-diffusion-2-1. [Online].

[4] *Stable Diffusion Github*. https://github.com/CompVis/stable-diffusion. [Online].

[5] *Using LoRA for Efficient Stable Diffusion Fine-Tuning*. https://huggingface.co/blog/lora. [Online].

[6] *Vintage dataset*. https://huggingface.co/datasets/Norod78/vintage-blip-captions. [Online].

[7] *Weights and Biases*. https://wandb.ai/. [Online].