CSC6051/MDS6004 - Image Processing and Computer Vision – Final Project
Enhancing Generative Art: Fine-Tuning Stable Diffusion 3 Medium with SimpleTuner

Zijin CAI: (40%)
224040002@link.cuhk.edu.cn

Shunuo SHI: (30%)
120090216@link.cuhk.edu.cn

Yuxuan ZHU: (30%)
120040089@link.cuhk.edu.cn

MSc in Data Science
School of Data Science

The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong Province, China

Abstract

This research aims to explore the potential of fine-tuning the Stable Diffusion 3 Medium (SD3M) model using the SimpleTuner toolkit to generate higher quality and more diverse 2D concept art. The project focus on understanding the technical differences between fine-tuning the SD1.5/SDXL and SD3M, and leveraging these insights to develop a robust fine-tuning methodology comprehensively that can be applied across various artistic domains using SimpleTuner. Further efforts are conducted on the evaluation of the SD3M models in generating concept arts, as well as insights and learnings from the fine-tuning process to support future research and applications.

1. Introduction and Literature Review

The generative art has been revolutionized by the advent of AI-driven diffusion models, and with Stable Diffusion being at the forefront. The latest SD3M, offers improved capabilities for generating detailed and stylistically consistent images. However, the full potential of these models can be unlocked through fine-tuning, which allows for customization and specialization to specific artistic needs. SimpleTuner is a versatile fine-tuning toolkit, provides a straightforward approach to above process. This research will investigate the effectiveness fine-tuning SD3M using SimpleTuner and document study outcomes in terms of artistic quality and diversity.

Diffusion models have emerged as the influential tools for deep generative models, with the capability performing in applications such as image synthesis, video generation, and the text-to-image translation designs. [Ho, Jonathan and Chan, William and others 2024] These models create data by reversing the process of diffusion that incrementally ascend the noise level of the data until it reaches the desired level (i.e., Gaussian distribution). The iterative procedure consists of training the model to predict the noise that has been added to the sample at each step, effectively denoising the generative image from the text-prompt. The approach can thus be contributed to significant advancements in generating detailed and coherent images, and allows the diffusion models to be applicable for the tasks that require high-fidelity generation such as text-to-image synthesis and image inpainting. [Md Manjurul Ahsan and Shivakumar Raman and Yingtao Liu and Zahed Siddique 2024]

The Stable Diffusion3 (SD3M) is one of the state-of-art diffusion model with 2 billion parameters, which is widely recognized for its efficient inference speed and excellent generation quality. SD3M has been open-sourced by OpenAI, providing a pre-trained model that can be fine-tuned to generate high-quality images in various artistic domains. It represents significant advancements in the field of diffusion models, particularly in the concept-art-image generation.

SimpleTuner is the fine-tuning toolkit simplifying the process of customizing and fine-tuning pre-trained models including SD3M. It allows for shifting the trained portion of timestep schedules using a simple decimal value, which can be used to adjust model's performance and generate outputs with diversity. The SimpleTuner also supports quantized model training, which helps to reduce the requiring precision and VRAM usage making it possible to train models, for instance Flux, on the consumer-grade GPUs (i.e., 16GB VRAM). [bghira 2023] This toolkit has been widely adopted in the research community for fine-tuning

diffusion models, and has been shown to be effective in enhancing the quality and diversity of generated images with efficiency.

Low Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique that leverages the low-rank structure of the model's weight matrices. It has emerged as a prevalent method without significantly appending the number of parameters, which is particularly useful for fine-tuning large-scale (generative) models with limited computational resources. [Zeng and Lee 2024]

The researches and repositories mentioned above have provided a solid foundation for this project, and the insights gained from these works will be instrumental in fine-tuning SD3M using SimpleTuner to generate high-quality and diverse concept art.

2. SD1.5/SDXL and SD3M

Model Architecture:

- SD1.5 and SDXL are the earlier versions of the Stable Diffusion model that, SD1.5 features a single generative model with 86 million parameters, while SDXL has 260 million parameters in total ($3\times$ than SD1.5), with the components of one base model for latent space up-scaling, and one refiner model for detailed up-scaling. [CSDN Community 2024]
- SD3M offers a range of model sizes from 800 million to 8 billion parameters with different volume versions, and introduces the Multimodal Diffusion Transformer (MMDiT) architecture [1] that significantly improves the model's semantic consistency capabilities.

Fine-Tuning Methods:

- Fine-tuning SD1.5/SDXL typically involves adjusting the short-text prompt and long-sentence prompt, while SDXL has improved inference for long-sentence that can be drawn from specific languages like English.
- SD3M supports more flexibility with full fine-tuning and LoRA training: [Stability AI 2024]

$$\mathbf{W}_{LoRA} = \mathbf{W}_{original} + \nabla\mathbf{W} \tag{1}$$

where $\nabla\mathbf{W}$ is the low-rank adaptation matrix.

Model Performance:

- SD1.5 has a default pixel resolution of 512×512, while increasing the resolution can result in distortion and image degradation. SDXL starts with the default 1024 pixels resolution offering finesse images with more details, but cost significantly increased GPU resources.

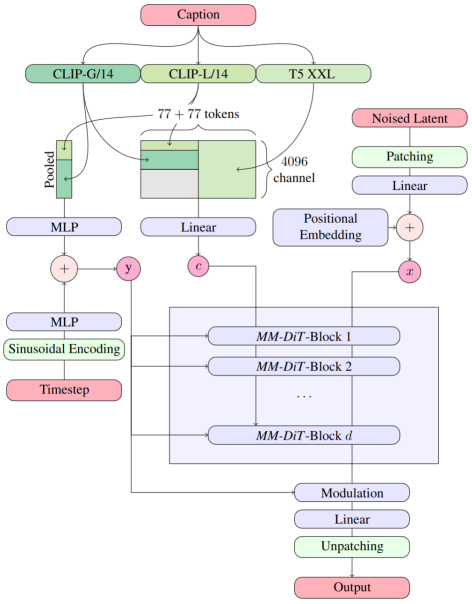

Figure 1. Multimodal Diffusion Transformer (MMDiT)

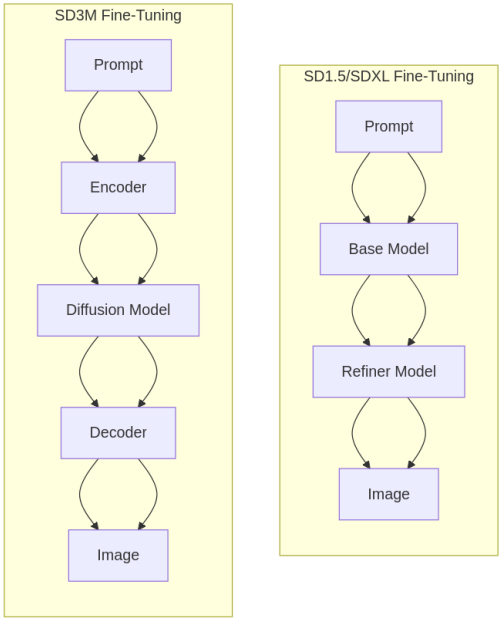

Figure 2. Flowcharts for SD1.5/SDXL and SD3M

- SD3M also begins with 1024 pixels resolution, but allows for more efficient inference and better generation quality compared to SDXL, particularly in the control of lights and colors. Nevertheless, SD3M is more efficient in terms of GPU memory usage, requiring less than 16GB VRAM to operate smoothly.

The simplified flowcharts 2 represent the fine-tuning process for SD1.5/SDXL and SD3M models.

3. SD3M Fine-Tuning: SimpleTuner

The guidelines provided by the SimpleTuner repository offer end-to-end instructions detailing fine-tuning process, including configure Microsoft's DeepSpeed for optimiser state offload for memory limited efficiency, and multi-node distributed training for large-scale models. [1]

The following part of this section provides some noticeable markdown for the SD3M implementation.

- `python-3.10/3.11` and `poetry install`:

There are built-in dependencies for SD3M (see the files `poetry.lock` and `pyproject.toml`), which are only supported by the specific python version. The following `tool.poetry.source` will accelerate the installation process in most cases:

```
name = "mirrors"
url = "https://to/your/mirror"
priority = "primary"
```

- Huggingface Hub and WandB:

Choose wisely for the network connection, try to `export` environment variables are suggested:

```
export HF_ENDPOINT=https://hf-mirror.com
export WANDB_API_KEY=your_api_key_here
```

- Configuration Files:

The experimental script `configure.py` provides interactive step-by-step configuration. However, the dataloader configuration is not included, thus the one should manually modify the `multidatabackend.json` file:

```
"cache_dir_vae": "cache/vae/sd3/xxx"
"instance_data_dir": "/path/to/your/data"
"cache_dir": "/cache/text/sd3/xxx"
```

- Quantized Model Training

Hugging Face Optimum-Quanto can be used to reduce the precision and VRAM requirements. There are several options for precision on `base_model` and `text_encoder`, as well as choices for different `optimizer`.

Please also be aware of the substantial dataset in fine-tuning process, that the dataset size need to be large enough to train the model effectively. However, there are additional memory usage on dateset encode-decode process, apart from loading the pre-train models. It is recommended to choose size-wise dataset base on the available hardware.

[1] Github Repo: https://github.com/bghira/SimpleTuner

4. SD3M Evaluation and Generative Art

The pre-trained SD3M model is fine-tuned with paintings by John Singer Sargent (Wiki-Art and auto-captioned). The following gallery 3 and 4 compare the baseline and the fine-tuned images in generating concept art.

example prompt: *a young woman with messy blonde hair and purple eyes, a slight smile, a pointy fantasy ear, a black feather hair tie, a pink feather, silver earrings, a white shirt, a black cloak, and a yellow rim light during golden hour.*

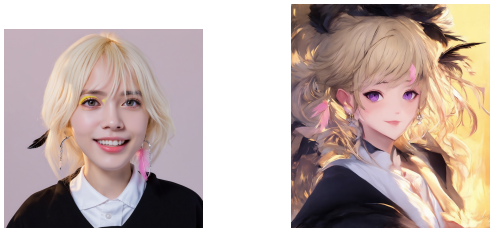

Figure 3. `base_model` (left) and `fine-tuned_model` (right)

example prompt: *A front wide view of a small cyberpunk city with futuristic skyscrapers with gold rooftops situated on the side of a cliff overlooking an ocean, day time view with green tones, some boats floating in the foreground on top of reflective orange water, large mechanical robot structure reaching high above the clouds in the far background, atmospheric perspective, teal sky.*

Figure 4. `base_model` (left) and `fine-tuned_model` (right)

An interesting observation is that, when performing inference for both `base_model` and `fine-tuned_model`, the suggestion is to adjust the prompt content to prevent `zzzq`. Especially for portrait-orientated text-to-image task, try less similarity to the prompt content from the internet, or simply vary the resolution pipeline of generated image.

More details please refer to the [2] `model_card` includes the training and validation settings, lycoris configuration, features of the dataset, and the step-by-step instruction of implementing the fine-tuned text-to-image directly.

[2] HuggingFace Model: https://huggingface.co/jimchoi/simpletuner-lora

Figure 5. Plots Summary of LoRA Fine-Tuning Process

The plots in 5 summarize the fine-tuned process that, the training loss shows fluctuations but generally trends downwards, indicating that the model is learning the features from the dataset. The learning rate demonstrates the linear decrease, which is consistent with the polynomial decay schedule used in the training. The gradient norm appears to be weill controlled which is mostly stable with only occasional spikes and the optimization process a good sign of general decreasing convergence.

5. Research Limitation and Documentation

- Dataset Size and Diversity:

The dataset fine-tuning the SD3M model in this research is relatively small. While the `Wiki-Sargent` dataset is sufficient to train the model effectively, it may not have been representative of the full range of artistic styles and concepts, potentially limiting the model generalizability.

- Computational Resources:

Consequentially, larger datasets are asking for more VRAM usage on image encoding-decoding, while the SD3M requires a significant amount of GPU memory to operate efficiently already The limited GPU resources constrained the batch size and the overall training speed, which may have impacted the training dynamics and may not fully utilize the capabilities (e.g., multi-GPU training environments).

- Fine-Tuning and Evaluation Techniques:

The report focuses on the use of SimpleTuner and LoRA for fine-tuning, while there are diverse methodologies available for fine-tuning diffusion models. For instance, models developed by the Stability AI also includes SD3-large and SD3.5M, etc. And the evaluation of concept art fine-tuned models is mainly based on visual inspection, more rigorous quantitative evaluation metrics could be conducted for more comprehensive assessment.

6. Conclusion

We appreciate the **CSC6051/MDS6004 - Image Processing and Computer Vision** course provides comprehensive understanding of the fundamental concepts and practical applications in the field. This course is certainly more valuable than any other courses we have taken.

We can tell from the report that, the workload is really intensive and time-consuming, but we admire the dedication and passion of Prof JIANG Li and all teaching assistants. Even the end of the semester is approaching, our group members have done our best to complete this report, and hopefully it meets the requirements of the course.

(mercy marking will be greatly appreciated, thank you!) :)

References

bghira (2023). *SimpleTuner/documentation/quickstart/FLUX.md at main · bghira/SimpleTuner*. https://github.com/bghira/SimpleTuner/blob/main/documentation/quickstart/FLUX.md. Accessed: 2024-12-08.

CSDN Community (2024). *Stable Diffusion 3*. CSDN. URL: https://blog.csdn.net/2401_84760527/article/details/140174463.

Ho, Jonathan and Chan, William and others (2024). *Diffusion Models: A Comprehensive Survey of Methods and Applications*. Tech. rep. DOI: 10.1145/3626235. New York, NY: Association for Computing Machinery.

Md Manjurul Ahsan and Shivakumar Raman and Yingtao Liu and Zahed Siddique (2024). *A Comprehensive Survey on Diffusion Models and Their Applications*. Tech. rep. arXiv:2408.10207v1. Norman, Oklahoma: University of Oklahoma.

Stability AI (2024). *Stable Diffusion 3 Medium Fine-tuning Tutorial*. Stability AI. URL: https://stability.ai/learning-hub/stable-diffusion-3-medium-fine-tuning-tutorial.

Zeng, Yuchen and Kangwook Lee (2024). "The Expressive Power of Low-Rank Adaptation". In: *arXiv preprint arXiv:2310.17513*.