

Stable diffusion customizing final-report

ByeongGeun Shin¹,
¹Pusan National University.

1 Stable diffusion Custom Fine-Tuning Practice

1.1 Introduction

딥러닝 프로그래밍 마지막 과제는 Stable Diffusion[1] 이미지 생성모델의 커스텀 파인튜닝을 진행해보는 것으로, Low-Rank Adaptation (LoRA)[2] PEFT 기법을 사용하여 Stable Diffusion 모델을 미세조정한 과정과 결과를 서술한다. 본 프로젝트의 주요 목표는 사용자 정의 데이터셋을 활용하여 특정 이미지 생성 작업에서 모델의 성능을 향상시키는 것이다. Stable Diffusion 모델의 미세조정은 데이터셋 준비, 이미지 캡셔닝, 메타데이터 생성, 모델 구성 및 학습, 추론 등의 여러 단계로 구성되어 있다.

1.2 Environment

실습 환경은 Python=3.11 Conda Env 가상환경에서 진행했고, 제공된 필요 패키지 버전을 맞춰야 했다. 다만 제공되는 패키지들을 그냥 깔면 의존성 문제가 발생하기에 다음과 같은 순서로 작업하였다.

```
!conda install pytorch==2.1.2 torchvision==0.16.2
  → torchaudio==2.1.2 pytorch-cuda=12.1 -c pytorch -c
  → nvidia #torch gpu
!pip install accelerate==0.26.1
!pip install datasets==2.16.1
!pip install diffusers==0.26.0
!pip install invisible-watermark==0.2.0
!pip install omegaconf==2.3.0
!pip install opencv-python
!pip install peft==0.7.1
!pip install scipy==1.12
!pip install transformers==4.37
!pip install bing-image-downloader
!pip install huggingface_hub==0.23.2
```

diffusers에서 모델을 다운로드 하기 위해 huggingface_hub 패키지가 자동으로 깔리게 되는데, 문제는 0.26.0에서 자동으로 깔리는 huggingface_hub는 최신버전이라. 현재 diffusers버전의 모델 다운에 필요한 함수가 deprecated되어 존재하지 않는다. 따라서 이후 다운그레이드를 해야 의존성 문제가 해결된다. 넘파이는 torch에 깔리는 기본 넘파이를 활용하고, 다른 버전을 깔면 또 다른 의존성 문제가 생기므로 건들지 않는다.

1.3 Dataset

1.3.1 Image Crawling

LoRA 학습의 첫 단계는 학습에 적합한 이미지 데이터셋을 수집하는 것이다. Bing의 공개 라이선스 이미지를 활용하여 다양한 주제와 스타일의 이미지를 수집하였다. 크롤링은 from bing_image_downloader import downloader 라이브러리를 이용해 쉽게 쿼리에 가까운 이미지들을 수집할 수 있다. 100개 정도를 목표로 "minecraft landscape"라는 쿼리로 마인크래프트 풍경 화면을 커스텀할 데이터셋 주제로 삼아 크롤링을 진행했고, 오류로 몇개가 건너뛰어지고 총 116개 이미지가 수집되었다.



Figure 1: "minecraft landscape" query image.

1.3.2 Image Captioning

수집한 이미지에 대해 Salesforce/blip-image-captioning-large[3] BLIP[4] 이미지 캡셔닝 모델을 사용하여 각 이미지에 대한 설명을 자동으로 생성하였다. 생성된 캡션은 이미지의 주요 특성을 반영하며, 모델이 텍스트와 이미지 간의 관계를 학습하는 데 활용되었다. 예를 들어, 특정 장면이나 스타일을 묘사하는 문구가 포함되도록 하였다.

1.3.3 MetaData

이미지와 캡션을 체계적으로 관리하기 위해 CSV 형식의 메타데이터 파일을 작성하였다. 각 항목은 이미지 파일 이름과 해당 캡션으로 구성되었으며, pandas 라이브러리를 사용하여 생성되었다. 데이터프레임을 활용하여 캡션과 파일 이름을 정리한 뒤, 이를 요구되는 형식에 맞춰 CSV로 저장하였다. 다만 그림 묘사에 마인크래프트를 인식할 때도 있고 아닐 때도 있으며 상당히 불안정해서 re모듈을 이용해 반복되는 글자들을 지우고, 묘사에 마인크래프트 태그가 없을 경우 맨 뒤에 "in minecraft" 문구를 추가해 좀 더 높은 연관성의 텍스트로 전처리하였다.

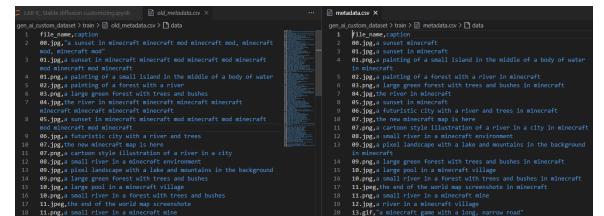


Figure 2: (a)Not preprocessed MetaData (b)Preprocessed MetaData .

1.4 Model

1.4.1 Use Pretrained Model and Setting

Stable Diffusion 모델의 "runwayml/stable-diffusion-v1-5"[5] 사전 학습 버전을 활용하였다. 학습 효율성을 높이고 메모리 사용량을 줄이기 위해 FP16 정밀도를 설정하였다.

1.4.2 Add LoRA Layer

모델에 LoRA 레이어를 추가하여 사전 학습된 가중치를 변경하지 않고도 작업별 특성을 학습할 수 있도록 구성하였다. 이는 계산 자원 절약과 특정 작업 적합성을 동시에 달성하는 데 기여하였다.

1.5 Training

1.5.1 Initialize

프롬프트 처리를 위한 토크나이저와 텍스트 인코더, 확산 모델링을 위한 노이즈 스케줄러, 그리고 Stable Diffusion 아키텍처의 주요 컴포넌트인 VAE와 UNet을 포함하여 학습 구성 요소를 설정하였다. LoRA 레이어를 제외한 나머지 매개변수는 requires_grad_(False)로 고정하여 학습 속도를 높였다.

1.5.2 Process

총 10,000단계의 학습이 진행되었다. 각 단계에서 모델은 입력된 캡션과 관련된 이미지를 생성하는 방식으로 학습하였다. 반복적인 학습 과정은 모델이 시각적, 문맥적 일관성을 갖춘 출력을 생성할 수 있도록 하였다. 훈련 도중 5000번 Step 중에서 Early Stop으로 훈련이 중단되었다. 시간은 약 2시간이 소요되었다.

1.6 Result

1.6.1 Inference

다시 runwayml/stable-diffusion-v1-5 모델을 파이프라인으로 로드하고, load_lora_ 함수를 이용해 이전에 훈련한 LoRA 어댑터 가중치를 로드한다. 이후 프롬프트로 "minecraft landscape of city view" 를 주어 이미지를 생성하고 테스트해보았다.



Figure 3: "minecraft landscape of city view" Prompt.

목표한 그림이 프롬프트에 맞게 잘 생성된 것을 볼 수 있었다.

References

- [1] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *arXiv preprint arXiv:2112.10752* (2022).
- [2] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *arXiv preprint arXiv:2106.09685* (2021).
- [3] Salesforce. *BLIP Image Captioning Large*. Accessed: 2024-12-23. 2022. URL: <https://huggingface.co/Salesforce/blip-image-captioning-large>.
- [4] Junnan Li et al. “BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *arXiv preprint arXiv:2201.12086* (2022).
- [5] RunwayML. *Stable Diffusion v1-5*. Accessed: 2024-12-23. 2022. URL: <https://huggingface.co/runwayml/stable-diffusion-v1-5>.