Use of Generative Adversarial Networks (GANs) and LoRA Stable Diffusion for Image Generation

Matthew Jamison, James A. Nguyen, Lawrence Mei
University of California, Irvine
Donald Bren Hall, 6210, Irvine, CA 92697
mjamiso1@uci.edu, janguye5@uci.edu, lawrenm1@uci.edu

Abstract

This project aims to generate images inspired by the Japanese Ukiyo-e Edo period art style, using approximately 1,500 artworks as references. By replicating historical styles, the project can transform modern subjects into 17th-century Japanese art. Applications include art restoration, concept art creation, and humorous artwork.

1 Introduction

The purpose of our project is to generate new images inspired by a specific art style, using approximately 1,500 Japanese Ukiyo-e Edo period art pieces as our reference. Achieving this allows us to create new art or images that replicate historical styles, such as turning modern objects or people, like Professor Baldi, into 17th-century Japanese art. This approach has several applications, including art restoration, creating concept art, or simply producing amusing art pieces. Additionally, it highlights the importance of discerning between AI-generated and real images, particularly for internet safety. In a world where fake news and fabricated images circulate widely, our project underscores the need for tools that can differentiate between genuine and AI-created content, which is crucial for maintaining trust and accuracy in digital media.

2 Data

To create a dataset suitable for training our model to mimic Ukiyo-e Edo period art, we began with a large Kaggle dataset comprising 30,000 art pieces from various cultures. Our goal was to isolate and focus on Japanese art, specifically from the Ukiyo-e Edo period. Initially, we manually identified and isolated the images corresponding to Japanese art, reducing our dataset to 2,236 images. We further filtered these images to include only those representing the Ukiyo-e Edo period style and were of high enough resolution. This gave us a total of 1,492 images to work with for both models. To ensure consistency in training, we standardized all images to a resolution of 480x320 pixels, excluding those that could not meet this criterion. For the LoRA training, we downsized images to 600x400 because of access to better hardware and faster training times. Precise labels for each image were generated using BLIP captioning, followed by manual correction to ensure accuracy. This meticulous process resulted in a final dataset, standardized in resolution and accurately labeled, providing a robust foundation for our model training.

3 Methods and Implementation

Our attempt to replicate the art style consistent throughout our dataset involved experimenting with a generative adversarial networks (GAN) model and building upon a pre-trained low-rank adaptation (LORA) model.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3.1 GAN Generator

The generator network of the GAN showed slight improvement when trained with the RMSprop optimizer compared to the Adam optimizer. This network included a dense layer architecture, where each successive layer decreased in size to facilitate hierarchical feature generation. Each hidden layer incorporated batch normalization to expedite and stabilize the training process, along with upsampling methods to double the spatial dimensions progressively. Starting from a 100-dimensional input noise vector, the generator produced a 480x320 output, suitable for input to the discriminator network alongside a real sample from the original image dataset.

3.2 GAN Discriminator

The discriminator network, in contrast, applied downsampling and dense layers of increasing size to systematically reduce the spatial dimensions, thereby enhancing hierarchical feature extraction. During the training process, both equal and distinct learning rates for the generator and discriminator were experimented with. However, mode collapse was encountered in the early stages of training when maintaining a consistently lower learning rate for the discriminator. Despite implementing an exponential decay learning rate scheduler, it is proposed that future iterations of this model would benefit from a more adaptive learning rate scheduler, designed to decrease the learning rate only during the latter stages of training.

3.3 GAN Hyperparameters

In general, memory and hardware constraints limited training to low experimental batch-size and total number of epochs. Ultimately, our latest model demonstrated the following hyperparameters: epochs = 500, batch size = 16, learning rate = 0.0001, decay steps = 1000, decay rate = 0.95, epsilon = 1e-07, additional epochs = 301

3.4 LoRA

For the LoRA training of the SDXL 1.0 model, we utilized Low-Rank Adaptation (LoRA), which fine-tunes large pre-trained models efficiently by freezing the original model weights and introducing new, trainable low-rank matrices. This approach made the training process faster and less resource-intensive.

The SDXL 1.0 model was chosen for its robust architecture, which includes an encoder/decoder, diffusion process, language transformer, U-Net, and cross-attention mechanisms. Trained on a vast dataset of 2.3 billion images, it was ideal for generating high-quality art in the desired style.

Implementing LoRA involved adding low-rank matrices to the attention layers, with only these parameters being updated during training while the rest of the model remained static. This not only sped up the training process but also produced a smaller, more manageable model size.

3.5 LoRA Hyperparameters

Key hyperparameters included using the Adafactor optimizer to save on memory, setting a learning rate of 1e-3, a network rank of 128, a batch size of 20, and running for 12 epochs. Adafactor, a memory-efficient variant of Adam, reduces memory usage by computing first and second moments on a per-row or per-column basis. The relatively high learning rate of 1e-3 was chosen to allow for faster convergence due to the reduced number of trainable parameters. The network rank of 128 balanced training efficiency and model performance, while the batch size of 20 was a practical choice for computational load. Running for 12 epochs ensured thorough comparison and evaluation of model performance

4 Results and Analysis

4.1 GAN Loss Graphs

Although our GAN, unfortunately, reached hardware limitations at 439 completed epochs, the model achieved promising early-stage training results that highlight the development of simple feature

Figure 1: GAN Results Epoch 439

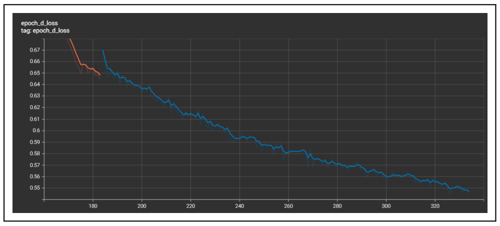

Figure 2: Discriminator Loss vs. Epoch

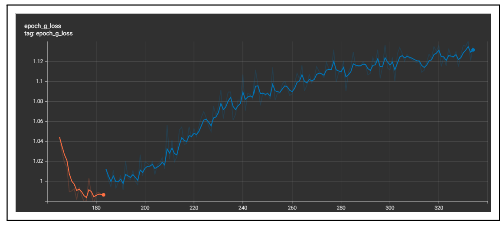

Figure 3: Generator Loss vs. Epoch

generation and color similarity to the original dataset(Figure 1). Trends observed in the generator and discriminator loss graphs suggest that the generator model fails to rival the discriminator around 180 epochs, which diminished the rate of improvement (Figure 2,3).

Overall, the GAN performed about as expected, given resource and time constraints, yet obvious improvements could be implemented to maximize these resources. Future additions to the model include the potential use of a Wasserstein loss function and the analysis of Frechet Inception Distance.

4.2 LoRA Outputs

We evaluated our LoRA-trained models by generating images with specific prompts using the same starting noise across all 12 models, selecting the top three from epochs 1, 6, and 7. The subjective nature of art made it challenging to determine which model performed "best."

Comparing outputs from these models with the base SDXL model showed significant stylistic differences, demonstrating the effectiveness of our training. The LoRA models successfully captured the Ukiyo-e Edo period style.

Figure 4: LoRA Outputs Compared by Epoch

Figure 5: LoRA - Godzilla Prompt

We also tested the models with pop culture references, such as Darth Vader, Gundams, and Godzilla. These references leveraged the pretrained SDXL model's extensive knowledge, allowing the LoRA models to adapt modern elements into the traditional art style seamlessly.

Interestingly, the model trained for just one epoch performed exceptionally well. This can be attributed to the strong foundation of the pretrained SDXL model, which allowed the initial epoch to imprint the broad stylistic features effectively. As training continued, overfitting to specific details may have reduced versatility, suggesting that early stopping could benefit artistic style training.

Overall, our results confirm the effectiveness of using LoRA for style transfer in stable diffusion models. The significant stylistic differences and successful integration of modern elements demonstrate the practical benefits and versatility of our approach

5 Conclusion and Discussion

Our project aimed to generate new images inspired by the Japanese Ukiyo-e Edo period art style, demonstrating the potential to blend historical styles with modern subjects for applications in art restoration, concept art creation, and humorous artwork. Additionally, our work highlights the importance of discerning AI-generated images from real ones to address internet safety concerns.

We curated a robust dataset of 1,492 high-resolution Ukiyo-e images, standardized for consistent training. Our methodology involved experimenting with a GAN model and enhancing it with a Low-Rank Adaptation (LoRA) model. The GAN's generator network, optimized with RMSprop, showed marginal improvements, while the discriminator faced early mode collapse, suggesting the need for a more adaptive learning rate scheduler in future iterations. The LoRA approach, using the SDXL 1.0 model, proved efficient and resource-saving, producing high-quality art with finely-tuned hyperparameters.

Despite hardware limitations, our GAN model demonstrated promising early-stage developments. Future improvements could involve integrating a Wasserstein loss function and analyzing Frechet Inception Distance to better evaluate performance. The LoRA-trained models, evaluated across 12 epochs, successfully captured the Ukiyo-e style and seamlessly integrated modern elements, such as pop culture references, demonstrating versatility and practical benefits. Interestingly, models trained for just one epoch performed exceptionally well, suggesting that early stopping might prevent overfitting and maintain versatility in artistic style training.

In conclusion, our project confirmed the efficacy of using LoRA for style transfer in stable diffusion models. The significant stylistic differences and successful adaptation of modern elements into traditional art styles illustrate the practical applications and future potential of our methods. Future developments will focus on refining training techniques, expanding datasets, and further exploring the intersection of AI and art.

6 Individual Contributions

6.1 Matthew Jamison

In this project, I contributed by helping clean and prepare the data, which involved curating and standardizing a dataset of Ukiyo-e Edo period art images. I also helped design the initial layout of our GAN, ensuring its structure was suitable for effectively learning the target art style. My main contribution was training and testing the LoRA model, selecting appropriate hyperparameters and running the training on RunPod's cloud GPUs. My responsibilities included evaluating model performance and comparing outputs to select the best-performing models.

6.2 James A. Nguyen

My contributions to the project centered around the refinement, hypertuning, and active evaluation of the generative adversarial network (GAN) model. This involved carefully analyzing live training to detect issues in the model such as mode collapse. This process was very tricky, as balancing the discriminator and generator model often failed in later stages while trying to correct early stage pacing inconsistencies, and vice versa. Doing so allowed the team to make the proper adjustments to the model architecture to assist in convergence towards a more successful model. Some additions to the model not yet implemented due to time and resource constraints include using a Wasserstein loss function and analysis of other loss metrics such as Frechet Inception Distance to help with hyperparameter tuning.

6.3 Lawrence Mei

I contributed to the project by running GAN's on the initial Takashi Murakami art and pokemon sprites. I also helped with the data preprocessing and looking at the individual images in each of the datasets to check and see if they aligned with our goals in the project. Furthermore, I verified that the datasets did indeed have the images that we wanted and manually removed blank or corrupt samples. Before switching to the ukiyo-e-edo art, I was also adjusting and testing various hyperparameters for the GAN to see if we could avoid early mode collapse and running epochs for the network to judge its performance.

References

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, W. (2021) LoRA: Low-Rank Adaptation of Large Language Models. arXiv. Retrieved from https://arxiv.org/abs/2106.09685.

Sayak, P. (2022) Using LoRA for Efficient Stable Diffusion Fine-Tuning. Hugging Face. Retrieved from https://huggingface.co/blog/lora.

bmaltais. (2023) Low-Rank Adaptation of Stable Diffusion Models. GitHub Repository. Retrieved from https://github.com/bmaltais/kohya$_s s$.

MachineLearningMastery.com. (2022) Using LoRA in Stable Diffusion. Retrieved from https://machinelearningmastery.com/lora-stable-diffusion/.

Paperspace Blog. (2022) Training a LoRA model for Stable Diffusion XL. Retrieved from https://blog.paperspace.com/lora-stable-diffusion-xl/.