Gotta Create 'Em All: Generating Pokémon with Stable Diffusion

Matej Miočić[1]

Abstract

Generative models have made significant advancements in generating realistic images across various domains. Pokémon, as a popular and visually diverse franchise, presents a unique challenge for generative models due to the complex details and specific art style associated with the Pokémon universe. This project presents an approach to fine-tuning stable diffusion for the generation of new Pokémon images.

Keywords

Stable Diffusion, Fine-Tuning, Pokémon

[1] *mm9520@fri.uni-lj.si, 63180206*

Introduction

Generative models have witnessed remarkable advancements in recent years, enabling the synthesis of images, text, and even music. Among these models, stable diffusion has emerged as a powerful technique for generating high-quality images by gradually transforming a simple distribution into a target distribution. By leveraging this approach, we aim to capture the essence of Pokémon species and faithfully recreate their visual characteristics in generated images.

However, generating realistic Pokémon images poses unique challenges. Pokémon designs encompass a wide range of shapes, colors, textures, and complex details, requiring a generative model to capture both the macro-level structure and the micro-level nuances of these creatures.

To address these challenges, we propose fine-tuning stable diffusion, a technique that builds upon the foundations of stable diffusion while incorporating Pokémon-specific data. Our approach involves initializing a pre-trained stable diffusion model and fine-tuning it using a dataset of Pokémon images, allowing it to learn the underlying distribution of Pokémon visual features.

Similar work was already done by Justin Pinkney at Lambda Labs [1], where they used 833 BLIP captioned Pokémon images [1] for fine-tuning a stable diffusion model and achieved incredible results.

The objective of this project is to explore whether more comprehensive image captions obtained from Pokémon Bulbapedia [2] and additional enhancements addressed in the following section can generate superior results.

[1] https://huggingface.co/lambdalabs/sd-pokemon-diffusers
[2] https://bulbapedia.bulbagarden.net/wiki/Main_Page

Methods

In this section, we will go through our data preparation which involves web scraping and some minor manual fixes. Then we will introduce the basics of stable diffusion and the process of fine-tuning.

Data preparation

We prioritize improving the captions associated with BLIP captioned Pokémon images [1] as our primary objective. The approach employed in their study involved utilizing a pre-trained BLIP model [2] to generate captions for each image. However, we have observed that these captions can occasionally be misleading.

This is why we made the decision to extract visual descriptions of Pokémon from Pokémon Bulbapedia through web scraping. These descriptions are often long so we experimented with variations of one, three, and all sentences. We also experimented with including Pokémon type and Pokémon evolution stage at the beginning of the sentence. Pokémon types were obtained from Kaggle dataset[3] and Pokémon evolution stages were obtained with a Python script using Pokémon API [4].

Another challenge encountered in BLIP captioned Pokémon images [1] was their tendency to overfit to specific Pokémon, particularly due to the presence of multiple images capturing slightly different forms of the same Pokémon. To address this issue, we made the decision to remove certain Pokémon from the original dataset, resulting in a final selection of 721 Pokémon images.

[3] https://www.kaggle.com/datasets/rounakbanik/pokemon
[4] https://pokeapi.co/docs/v2

In the end, we successfully created four distinct datasets, each consisting of the same set of images but accompanied by varying captions. For the first three models, we fine-tuned them using datasets that contained original descriptions from Pokémon Bulbapedia. These descriptions ranged in length, comprising of all, 3, and 1 sentence(s) respectively. Additionally, we included the Pokémon type and evolution stage at the beginning of the first sentence in these datasets. In the case of the third model, we excluded the Pokémon name from its dataset. Finally, the fourth model underwent fine-tuning using 1 sentence-long descriptions, without Pokémon's type, evolution stage, or name. The 1 sentence descriptions were also tuned manually. For simplicity we will call these four datasets **A**[5], **3S**[6], **1S**[7] and **1SN**[8] for the rest of this report.

Table 1. Table of captions for a Pokémon named Cyndaquil for different datasets.

BLIP	a cartoon bird with a hat on its head
A	Basic fire Cyndaquil is a small, bipedal mammalian Pokemon with bluish fur on top of its body and cream-colored fur on its underside. Its eyes are usually seen closed, and it has a long, thin snout. Its arms are short, but its legs are slightly more developed and have a single nail on each foot. On its back are four red circles that can erupt into flames. When the flames are burning, its back has the appearance of being covered in spines. Cyndaquil is capable of defending itself using these flames, and the fire grows more powerful as it becomes angry or defensive. However, if it is tired, the flames are not able to burn properly. It is usually timid by nature and often curls into a ball when intimidated. Though rare in the wild, it can be found living on grasslands.
3S	Basic fire Cyndaquil is a small, bipedal mammalian Pokémon with bluish fur on top of its body and cream-colored fur on its underside. Its eyes are usually seen closed, and it has a long, thin snout. Its arms are short, but its legs are slightly more developed and have a single nail on each foot
1S	Basic fire small, two-legged mammalian with bluish fur on top of its body and cream-colored fur on its underside.
1SN	small, two legs mammalian with bluish fur on top of its body and cream-colored fur on its underside

In Table 1, it becomes evident that BLIP captions occasionally fall short in accurately depicting the images. In contrast, the descriptions obtained from Pokémon Bulbapedia prove to be significantly more informative and visually accurate (see Figure 1). Additionally, it is worth noting that the descriptions in the table tend to be excessively long, so we experimented with shorter lengths.

Figure 1. An image of a basic stage and fire-type Pokémon named Cyndaquil.

Stable diffusion

Diffusion models are a relatively recent addition to a group of algorithms known as generative models. The goal of generative modeling is to learn to generate data, such as images or audio, given several training examples. A good generative model will create a diverse set of outputs that resemble the training data without being exact copies.

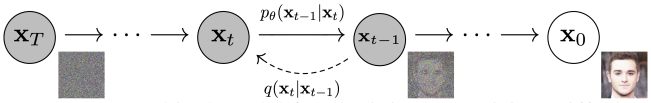

Figure 2. Graphical model from original Denoising Diffusion Probabilistic Models paper. [3]

The secret to diffusion models' success is the iterative nature of the diffusion process (Figure 2). Generation begins with random noise, but this is gradually refined over several steps until an output image emerges. At each step, the model estimates how we could go from the current input to a completely denoised version. However, since we only make a small change at every step, any errors in this estimate at the early stages (where predicting the final output is extremely difficult) can be corrected in later updates.

To generate new images with a trained model, we begin with a completely random input and repeatedly feed it through the model, updating it each time by a small amount based on the model prediction.

Fine-tuning

Fine-tuning is a common practice of taking a model which has been trained on a wide and diverse dataset, and then training it a bit more on the dataset you are specifically interested in. This is common practice in deep learning and has been shown to be tremendously effective.

Fine-tuning offers several advantages. It reduces the need for training deep learning models from scratch, which can be computationally expensive and time-consuming. By leveraging pre-trained models, we can save significant resources

[5] https://huggingface.co/datasets/matemato/pokemon_bulbapedia_all
[6] https://huggingface.co/datasets/matemato/pokemon_bulbapedia_3_sentence
[7] https://huggingface.co/datasets/matemato/pokemon_bulbapedia_1_sentence
[8] https://huggingface.co/datasets/matemato/pokemon_bulbapedia_desc_only

and benefit from the knowledge acquired during the initial training. Additionally, fine-tuning can help mitigate the issue of limited labeled data, as it allows the model to leverage the general knowledge it gained from the pre-training phase.

In our scenario, our objective is to create images that capture the unique art style of Pokémon. However, we also want to retain the valuable knowledge from our pre-trained network, enabling us to generate images with fresh and varied Pokémon designs. It's important to find a balance during the training process, as allowing our model to train for extended periods runs the risk of it losing crucial information and becoming excessively fixated on our specific Pokémon art style.

We decided to use *stable-diffusion-v1-5-pruned.ckpt* [4] as our pre-trained model weights available on Hugging Face. We used the same model[9] for fine-tuning as BLIP captioned work [1].

We used NVIDIA Quadro RTX 8000 (48 GB VRAM) for fine-tuning which took about 7 minutes for an epoch and NVIDIA Quadro RTX 5000 (16 GB VRAM) for inference which takes about 1 minute for creating an image with 200 sampling steps.

Results

In this section, we present the results of our work. Firstly, we show the results obtained from different checkpoints of the model weights. Secondly, we demonstrate how incorporating additional information about the Pokémon's stage during training enables us to generate an entire evolution chain, including the baby form, basic form, and stage 1, using a single seed. Furthermore, we examine whether the model is capable of recognizing Pokémon types. Finally, we showcase several examples with varying levels of unconditional guidance scales.

Model through training

Figure 3 showcases a collection of Pokémon generated using different model checkpoints. Each row presents examples generated from a different model. The first row corresponds to a model trained on all sentences (**A**), while the second row represents a model trained on three sentences (**3S**). The third row displays examples from a model trained on one sentence (**1S**), and the fourth row illustrates results from a model trained on one sentence without additional Pokémon features (**1SN**) as mentioned in data preparation.

For the first three rows, we used checkpoints ranging from 11 to 99. As for the fourth row, we present checkpoints from 11 to 187, while omitting every other checkpoint. This is because we set a learning rate to $5.0e^{-05}$ for the last model instead of $1.0e^{-04}$. For the third and fourth rows we also used 10000 warm-up steps which is observable in the first column.

We can observe slight variations among the generated images, which can be attributed to the changes in model parameters. Nonetheless, all the images maintain a notable resemblance to the provided prompt. This indicates that the model has successfully learned and retained the Pokémon art style, even with different prompts used in training.

[9] https://github.com/justinpinkney/stable-diffusion

Figure 3. Images generated with 4 models (different model in each row) through training (different checkpoints in each column) with same seed for a prompt: "A graceful snow leopard, with a sleek, snow-white coat dotted with dark, rosette-like markings.".

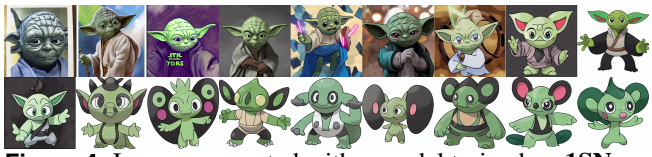

Figure 4. Images generated with a model trained on **1SN** dataset through training for a prompt: "Yoda".

In Figure 4, we demonstrate the process of training the model, where it gradually converges towards capturing the art style of Pokémon. However, as the training duration extends the model tends to forget certain aspects, resulting in a final output that deviates from the original prompt. We used checkpoints from 11 to 187 resulting in 18 images.

In the appendix, we present a more extensive collection of Pokémon image examples, featuring larger images for better visibility and detail.

Evolution chain

In the Pokémon universe, Pokémon go through physical transformations known as evolution, where they often retain their type and certain visual characteristics. We conducted experiments by adding the evolution stage information to our training prompts. By including this additional detail, we aimed to enhance the model's ability to generate Pokémon designs that accurately reflect the progression from one stage to another. This approach allowed us to explore the generation of complete evolution chains, encompassing the baby form, basic form, stage 1, and stage 2 while maintaining the essential characteristics associated with each Pokémon type.

In Figure 5, we show that the first three models are able to generate coherent evolution chains. Each row in the figure represents a different model. The generated examples demonstrate the model's capability to create a sensible progression from the baby form to stage 1 of Pokémon evolution.

In the last row of the figure, we feature a model trained on data without evolution stage information. As a result, the

Figure 5. Images generated with 4 models (different model in each row) with same seed for prompts: "Baby Fire Bird.", "Basic Fire Bird." and "Stage 1 Fire Bird." in each column.

distinctions between the basic form and stage 1 may not be as noticeable.

Pokémon types

Pokémon types are an essential aspect of the franchise. Each type represents a unique elemental attribute, ranging from fire and water to grass and electric, among many others. These types also influence Pokémon's appearances. Fire-type Pokémon often exhibit fiery features, such as flaming manes or tails, while water types tend to possess aquatic characteristics like fins or streamlined bodies. Grass types showcase vibrant foliage, with leaves or vines adorning their forms.

We added the Pokémon type to our prompts to see if our models could learn to differentiate type appearances.

Figure 6. Images generated with 4 models (different model in each row) with same seed for prompts: "Baby Water Bird.", "Baby Grass Bird.", "Baby Dragon Bird.", "Baby Fairy Bird." and "Baby Fighting Bird." in each column.

In Figure 6, we show that despite the fourth model being trained without any explicit type information, simply knowing the name of the type is sufficient to accurately represent the intended visual representation.

Unconditional guidance scale

The unconditional guidance scale plays a crucial role in text-to-image synthesis within the stable diffusion framework. It serves as a control mechanism that allows users to manipulate the level of influence by the textual guidance on the generated images. By adjusting the unconditional guidance scale, users can fine-tune the balance between the input text and the creative freedom of the model. A higher scale value strengthens the influence of the textual input, resulting in images that closely align with the given description. On the other hand, a lower scale value allows the model to exercise more artistic liberty, producing visually diverse interpretations that may deviate from the original text.

Figure 7. Images generated with model trained on **1SN** dataset for different guidance scales (5, 7.5, 10, 12.5, 15) for prompts: "Barack Obama", "Donald Trump", "Sonic", "Mario" and "Thor".

In Figure 7, we illustrate the influence of the unconditional guidance scale on image generation. As the parameter increases, the resulting image progressively aligns with the prompt but with a slightly artificial appearance.

Conclusion

In conclusion, this project has presented an approach to fine-tuning stable diffusion for the generation of new Pokémon images. By incorporating Pokémon-specific data and utilizing the stable diffusion framework, we have achieved significant progress in generating high-quality and visually consistent Pokémon designs.

Through the process of fine-tuning, we have successfully captured the essence of Pokémon art style and generated Pokémon images that closely resemble the provided prompts. The model has also demonstrated its ability to generate coherent evolution chains, showcasing the visual progression from one stage to another.

When comparing our work to BLIP captioned dataset we cannot say much about improvement due to the absence of an unbiased metric to evaluate our work. We did however show that captions do matter when fine-tuning to a specific dataset.

References

[1] Justin N. M. Pinkney. Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Appendix

In this section we provide some new and some already seen Pokémon image examples, featuring larger images for better visibility and detail. We also provide prompts for all images.

Images through training

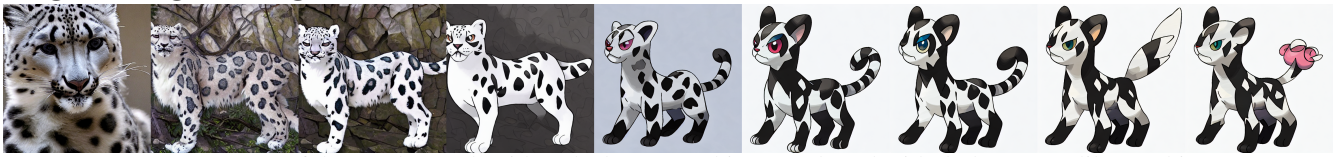

Figure 8. A graceful snow leopard, with a sleek, snow-white coat dotted with dark, rosette-like markings.

Figure 9. A majestic, ethereal butterfly, with delicate, translucent wings that shimmer with radiant hues.

Figure 10. A mystical, floating lotus flower, with petals that emit a soft, gentle glow.

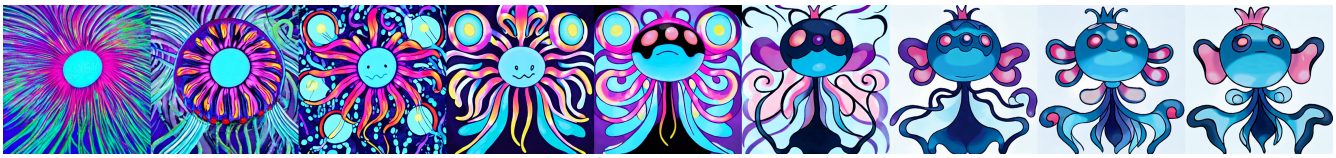

Figure 11. A bioluminescent jellyfish, glowing in a myriad of mesmerizing colors.

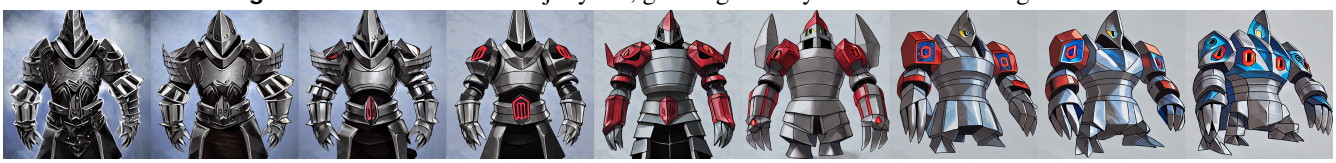

Figure 12. A fearsome, armored knight, with gleaming metallic plates, sharp edges, and glowing eyes.

Figure 13. A majestic peacock, with vibrant, iridescent feathers.

Figure 14. A mythical phoenix, with resplendent feathers that burn with brilliant flames.

Figure 15. A majestic, towering tree with branches that stretch high into the sky, adorned with leaves.

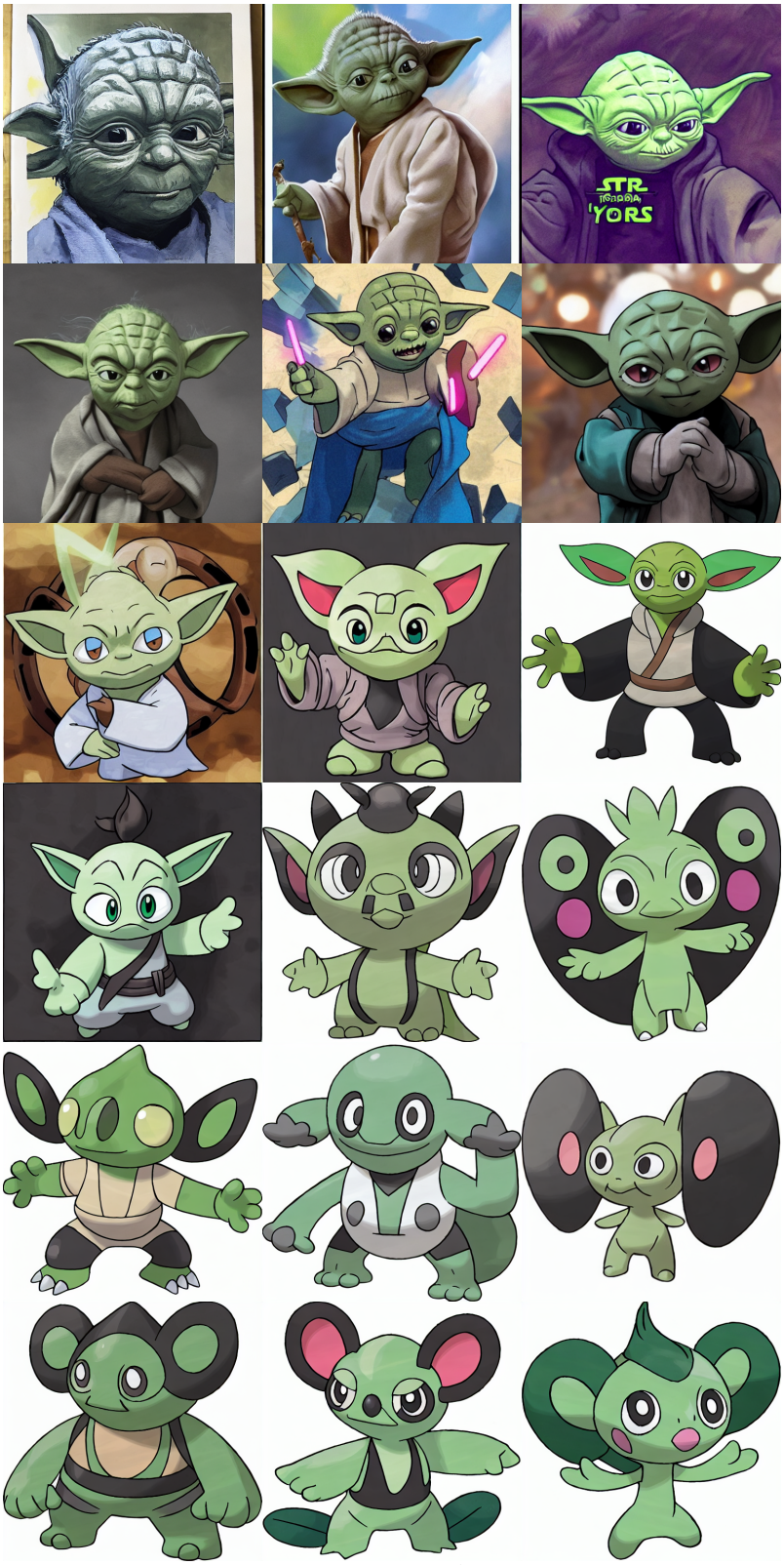

Figure 16. Yoda.

Evolution chain

Figure 17. Baby Fire Bird. Basic Fire Bird. Stage 1 Fire Bird.

Guidance scale (5, 7.5, 10, 12.5, 15)

Figure 18. Barack Obama.

Figure 19. Donald Trump.

Figure 20. Sonic.

Figure 21. Mario.

Figure 22. Thor.