Stable Diffusion Model Collapse

Gavin Pu

1 Introduction

Model collapse is a phenomenon defined by successive generations of generative models displaying degraded performance when the training set of one generation is derived from the outputs of the previous generation. Model collapse mainly occurs due to only being able to take a finite number of outputs using the model's underlying distribution at each generation, leading to potential information loss since the model distribution of the next generation is only an approximation of the model distribution of the current generation. Intuitively, outputs that are more probable tend to be oversampled, while outputs that are less probable tend to be undersampled. This causes the tails of the distribution to shrink over time and the distribution's variance to approach zero as successive generations of models are trained [4].

In gaussian mixture models (GMMs), model collapse was observed by [4] when fitting two Gaussians to a set of points in two-dimensional space and using the fitted Gaussians to generate a new set of points used for fitting the next generation of two Gaussians. While the model was observed to fit the data well during the initial generation, both Gaussians eventually collapsed to a state where they had very little variance by generation 2000.

Variational autoencoders (VAEs) were another kind of simple generative model studied by [4]. The MNIST dataset was used to train the original VAE, which appeared to then generate plausible handwritten digits [1]. However, by generation 20, all generated outputs appeared quite similar and resembled a mix of all digits. This suggests that the latent distribution of the VAE became more unimodal over time as modes from the distribution of the original VAE were merged together.

Unlike GMMs and VAEs which can be small enough to train from scratch, large language models (LLMs) are usually fine-tuned from a pre-trained model. [4] used the OPT-125m language model developed by Meta and fine-tuned the model on the WikiText dataset [2, 5]. Five separate model instances were fine-tuned on a training set for each generation, and the best model was selected using the WikiText validation set to serve as the base model to fine-tune for the next generation. The authors performed this experiment with two different configurations. In the first one, training datasets for the next generation were comprised entirely of data generated from the best model of the previous generation, and models were trained for five epochs. In the second, 10% of the data within training datasets were randomly sampled from the original WikiText training set while the remaining 90% were still generated using the best model of the previous generation, and models were trained for ten epochs [4].

For both configurations, perplexities of each individual training data sequence produced by models from all generations were calculated using the model from the first generation trained only on the original training set. The authors found that in later generations, models both were even more likely to generate outputs that the first model was already likely to produce. At the same time, models from later generations also were more likely to generate outputs that would have almost surely never have been produced by the first model, suggesting that errors accumulated in generated outputs over time cause successive generations of models to drift from the original distribution of data.

A model that was not evaluated by [4] but was mentioned at the beginning of their abstract is Stable Diffusion. We hypothesize that because Stable Diffusion is another kind of generative model, model collapse will occur when training successive generations of Stable Diffusion models on outputs from previous generations.

2 Methods

We selected SD-Turbo, a Stable Diffusion model trained by Stability AI, to fine-tune [3]. To create a training set used to fine-tune the model that would become the first generation, we searched for photos under the "Student Life" category of the University of Toronto Digital Media Bank. While over 2400 photos were found, only 75 were available for download. The training set comprised these 75 photos and the `main` text caption for each one as shown in Table 1.

Prompt Name	Prompt Text
`main`	a photo capturing student life on campus at the University of Toronto
`similar`	a photo capturing student life on campus at the University of Waterloo
`different`	a wide shot of Santa Monica Beach

Table 1: Prompts used during training and image generation.

Figure 1: Sample images generated with each prompt using the models from the first and last generations.

We trained 20 generations of models using an A100 GPU on Google Colab. Aside from the first generation, which was trained on the original dataset, we generated 75 images using the `main` prompt at each generation to be used as the training set for the next generation and used the state of the model in the current generation as the base model for the next generation. Each model was trained for 50 epochs. Additionally, we prompted the model at each generation with the `similar` and `different` prompts shown in Table 1 to generate 25 images per prompt for use in evaluation on whether effects would be seen for prompts similar and different to the `main` prompt used for training.

3 Results

Based only on qualitative observation, the first model did not appear to be fine-tuned well on the training set of University of Toronto student life photos. However, after fine-tuning the model for 20 generations, the images generated appeared to have more noise than the images from the first model as seen in Figure 1.

Generated images had dimensions 512×512 and three channels for red, green, and blue. Hence, each image may be represented as a three-dimensional tensor with dimensions $3 \times 512 \times 512$ where entries are integers between 0 and 255, inclusive. Let $\mathbb{I} = \{0, 1, \ldots, 255\}^{3\times512\times512}$ be the set of all possible images that can be generated from a model at any generation. We define the distance metric $d : \mathbb{I} \to [0, \infty)$ by

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{512}\sum_{k=1}^{512}(\mathsf{X}_{i,j,k} - \mathsf{Y}_{i,j,k})^2}$$

as a three-dimensional generalization of the Frobenius norm for matrices. We calculated distances between each unique pair of images for all three prompts generated from the models in the first and last generations. The distributions of distances is shown in Figure 2. For all three prompts, the distribution of distances between images generated from the model in the last generation is shifted to the right compared to from the model in the first generation.

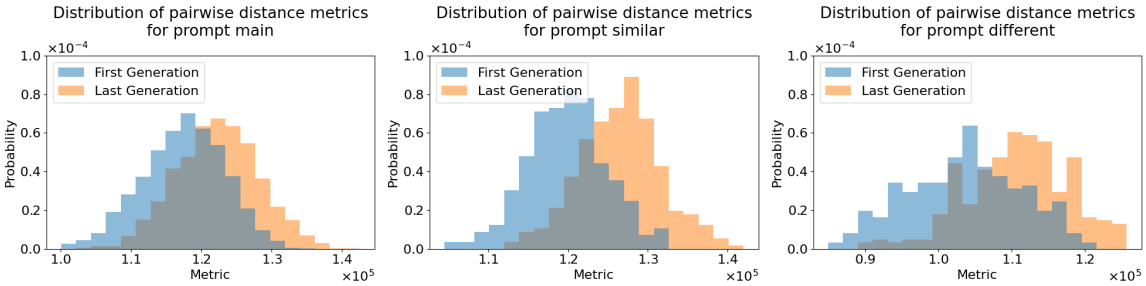

Figure 2: Distributions of pairwise distances for images generated with each prompt from the first and last generations.

4 Discussion

Against what would have been expected by model collapse, the last generation of SD-Turbo generated images that were more different from each other compared to the original fine-tuned model according to the distance metric defined previously. This held true across three different prompts, one of which was similar to the prompt used in training and another of which was more different. The images themselves, however, suggest that larger distances in images produced from the last generation could be caused in part by the model from the last generation appearing to produce more noisy output compared to the model from the first generation, though this was not quantified by a formal criterion.

The LLMs fine-tuned by [4] were found to have a greater chance of generating outputs very unlikely to be generated by the first LLM fine-tuned on the original data due to inaccuracies across successive generations causing errors to grow. While this may somewhat explain the increased distances between images when we fine-tuned Stable Diffusion models, we were unable to definitively observe model collapse. This could be because compared to the WikiText dataset used by [4], which contained tens of thousands of observations, our image dataset was extremely small [2]. Additionally, [4] took weeks to run experiments on fine-tuning LLMs while we did not have enough compute resources to fine-tune SD-Turbo for over a day. Hence, more training time may be necessary to observe model collapse in Stable Diffusion models.

Code Availability

The GitHub repository https://github.com/RobinEatingWorm/stable-diffusion-model-collapse contains all code used in and images generated from this experiment.

References

[1] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[2] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models, 2016.

[3] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation, 2023.

[4] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[5] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open Pre-trained Transformer Language Models, 2022.