MAT1510 Project

Diffusion Probabilistic Models Generalize when They Fail to Memorize

Boyang (Frank) Li

January 9, 2026

Summary of the Paper[1]

The paper delves into the training behavior of Diffusion Probabilistic Models (DPMs) through a set of hypotheses and their corresponding designed experiment. DPMs use two processes (forward diffusion process and reverse diffusion process) to learn how to draw images. During the first process, these models will destroy the input images by adding noise step by step, where they are trained progressively to learn the noise distribution. After that, by utilizing the learned parameters in the second process, these models will invert the noise addition and generate a data distribution by approximating the denoising step.

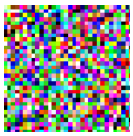
Utilizing the above training technique, DPMs are recognized for their superior performance in generating high-quality data distributions. However, concerns have arisen regarding their tendency to memorize training data, raising questions about both practical (privacy issues) and theoretical sides. To address this, the authors propose the *memorization-generalization dichotomy*, which means memorization and generalization are mutually exclusive phenomena in DPMs. This is in contrast to the current thought of supervised learning, where deep neural networks often exhibit "benign" overfitting—memorizing data while still generalizing effectively.

Hypotheses and Experiments

The study uses a series of controlled experiments to support this dichotomy. First, it defines memorization capacity as representing the maximum amount of data the model can memorize and finds that DPMs perform rote learning (memorization) up to that capacity. After that, the proportion of generated replication will decay rapidly when increasing the train set size. In addition, the study demonstrates that this dichotomy also manifests at the level of data classes. With the experiment on class-conditional DPMs with a train data set containing both majority (large population) and minority (small population) classes, they find that models tend to memorize minority classes while generalizing for majority ones.

A key aspect of the paper involves demonstrating that DPMs perform rote learning with sufficient training complexity but insufficient training size. To prove this, they add dummy data with disjoint data distribution with the original data into the train set to fill up the model capacity and then check whether the model performs conceptual learning (generalization). They do their first experiment on original data designed by simple procedurally generated data distribution called Mondriann (abbv. "MO") and dummy data called gaussian mixture (abbv. "GM") shown as the following,

(a) Mondrian



(b) Gaussian Mixture

With 2k images using MO distribution in the train set only, 92% of the generated images replicate the dataset. After that, adding 6k dummy data from GM into the train set, the trained model generates MO images conditionally where only 0.2% of them are replication. In addition, instead of MO images, the study picks 2k CIFAR-10 "car" images as the original train set, and augments the train set by adding $\{2, 4, 14\}$k GM images. Finally, the proportion of replicates that the trained model has generated decreases after adding more GM dummy images.

Limitation and Conclusion

However, the limitation still exists in their experiment. Especially as for the experiment of adding dummy data, GM may not provide a fully disjoint distribution compared to the original data, while the only dummy data that is used in the paper is GM. Therefore, the model may perform differently when using another type of dummy data.

Overall, the paper bridges theoretical insights with practical implications, demonstrating the ideas of memorization and generalization dichotomy. The findings lay a foundation for exploring how model capacity, data complexity, and training strategies collectively influence the performance of DPMs, and suggest that the learning dynamics of DPMs are fundamentally different from those of supervised models.

Motivation of My Experiment

My experiment focuses on the key aspect of the paper which illustrates when the train data complexity is sufficient but the size is insufficient, DPMs will perform rote learning instead of conceptual learning. Since GM is the only dummy data that the paper is using, then I would like to try different types of dummy data that add various information to the model and see how the model performs. Motivated by the paper[2], which illustrates that the diffusion generative models tend to learn the geometry-adaptive harmonic basis which is shaped by geometric

[1] https://openreview.net/forum?id=shciCbSk9h

[2] https://arxiv.org/abs/2310.02557

features of the image. My experiment will focus on adding dummy data of geometric shapes adaptive to the original data whose result will then be compared with GM dummy data.

Experiment

The experiment uses the model called IDDPM (Improved Denoising Diffusion Probabilistic Models)[3] with UNet-128 model. The model is trained from scratch using batch size 128 and with 100k iterations. As for the train data, it will start with containing 2k CIFAR-10 car images. After that, $\{1, 2, 4\}$k number of dummy data will be added into the train set to augment the set size. Dummy data considered in this experiment include GM image, GM and car combination, more CIFAR-10 car images, and some simple geometric elements (e.g. circle, grid, etc). Same as the GM used in the paper above, it is generated by (1) fixing 100 images with independent and uniform random pixel values within [0, 1]; (2) random selecting an image x from those 100 images and using the isotropic Gaussian distribution with mean x and standard deviation 0.5 for each pixel of the generated image; (3) clip the retrieved image by [0, 1]. In addition, GM and car combination image is generated using the idea of the hybrid image. That is, within the Fourier domain, combining 75% high frequency of a randomly generated GM image and 25% low frequency of a CIFAR-10 car image, such that the resulting image looks like a GM image but also preserves a car edges. Also, note that the CIFAR-10 car images used for generating hybrid images are chosen from the CIFAR-10 car dataset which is not in the original train set to avoid overfitting. Furthermore, since CIFAR-10 has 6k car images in total after combining the train and test sets, then the dummy data can also be chosen from the car images not in the original train set. Finally, a simple geometric element image is generated by randomly choosing a type containing circle, grid, gradient, and line, and then randomly setting the layout. The model is then trained by using all train sets generated above. Note that the model is trained unconditionally when the train set contains only car images, while it is trained using class conditional networks when the train set contains other types of images like GM or geometric elements.

(a) GM

(b) GM and Car

(c) More car images

(d) Circle

(e) Grid

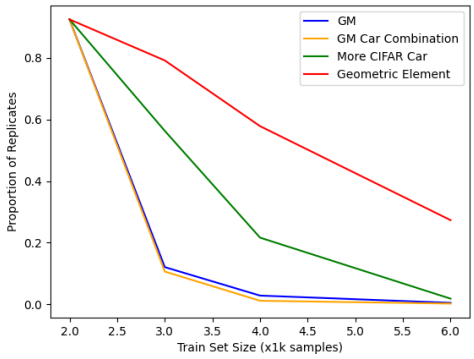
(f) Gradient

(g) Gradient

[3]https://github.com/openai/improved-diffusion

Evaluation and Result

After all models are trained, to evaluate the training result, all models will sample 4k car images, and the proportion of replicates will be measured by comparing samples with each image in the train set. Same as the replicate detection criterion above, including (1) given a generated image W; (2) compute L^2 norm between W and all images in the train set, and find $X_1, X_2 \in S$ with smallest and second smallest L^2 norm; (3) mark W as replicate if $\frac{\|W-X_1\|_2}{\|W-X_2\|_2} < \frac{1}{3}$. According to the resulting image below, adding GM dummy data has done a great job of motivating IDDPM to generalize new car images. GM and car combination datasets have made a little improvement compared to GM. Adding more CIFAR-10 car images can finally catch up with the above two. However, simple geometric elements do not work well on encouraging IDDPM to generalize.



Limitation and Conclusion

There still exist limitations in my experiment. Firstly, instead of testing on one model, I would need to accomplish a comparative test on different models and see whether adding GM dummy data does a great job in different sizes or types of DPMs. Additionally, the resolution of car image is low (32 * 32), so each data will contain less information which makes it look similar to GM. Consider training DPMs with more complex images may lead to different performances on rote and conceptual learning. Finally, the replicate detection criterion, which only considers two L^2 norm values(smallest and second smallest), may be unsuitable. That is, if the model is trying to generate a linear combination of two car images, those two norm values will be similar, and we will fail to detect the replicates in that situation. Overall, my experiment shows that only a train set size of 2k is not enough to encourage IDDPM to perform conceptual learning. However, the 2k CIFAR-10 car image set has enough complexity to motivate the model to generalize. That is because, after adding more dummy data GM with disjoint data distribution, the proportion of replicates sampled by the model drops quickly. After experimenting with more types of dummy data, combining car images with GM images in the Fourier domain works best on model conceptual learning.