Stable Diffusion and Imagen Can Reproduce Training Data Almost Perfectly

Originally published at https://arquivo.pt/wayback/20240203190037oe_/https://blog.metaphysic.ai/stable-diffusion-and-imagen-can-reproduce-training-data-almost-perfectly/

January 31, 2023

A new academic collaboration, including contributors from Google, has declared that generative image models are indeed capable of reproducing training data, bolstering current criticism and lawsuits that make this claim.

In the new paper, titled *Extracting Training Data from Diffusion Models*, researchers from Google, DeepMind, ETH Zurich, Princeton, and UC Berkeley, perform a series of resource-intensive experiments that successfully 'extract' training images to within a reasonable tolerance of what anyone might consider to be a reproduction of the original image.

One of a limited number of specific examples provided for the new paper, illustrating the capabilities of generative image synthesis models to produce carbon copies of the data on which they were trained - though the researchers have deliberately withheld many other similar instances, they claim. Source: https://arxiv.org/pdf/2301.13188.pdf

However, chary of aggravating copyright concerns on these issues further, most of the other instances are described, and further examples limited to the researchers' ancillary experiments on training diffusion models.

In the above example, we see on the left an image from the LAION subset that was used to train Stable Diffusion; on the right, essentially the same image elicited from Stable Diffusion via a round of iterative prompting.

Noting that images included in training sets scraped from public data are not necessarily allowed to be used as training material, the paper only shows examples, as with American Protestant evangelist Ann Graham Lotz above, where the original license allowed for free distribution.

The paper states:

'With a generate-and-filter pipeline, we extract over a thousand training examples from state-of-the-art models, ranging from photographs of individual people to trademarked company logos. We also train hundreds of diffusion models in various settings to analyze how different modeling and data decisions affect privacy.

'Overall, our results show that diffusion models are much less private than prior generative models such as GANs, and that mitigating these vulnerabilities may require new advances in privacy-preserving training.'

Diffusion Models Less Private than GANs

The experiments conducted by the collaborators found that latent diffusion models such as Stable Diffusion leak twice as much potentially private information as Generative Adversarial Networks (GANs). This is because, the authors assert, diffusion models effectively use memorization as a reproduction strategy, whereas GANs use an adversarial system wherein a *discriminator* grades the system's efforts to reconstruct training data without ever explicitly showing it the data.

'As a result,' the paper observes, 'GANs differ from diffusion models in that their generators are only trained using indirect information about the training data (i.e., using gradients from the discriminator) because they never receive training data as input, whereas diffusion models are explicitly trained to reconstruct the training set.'

The study has caused the researchers to reconsider the currently-accepted definition of *overfitting* – the process by which a generative AI is disposed, through quirks of training or shortcomings in the training data, to reproduce the original trained images instead of learning 'guiding principles' from them.

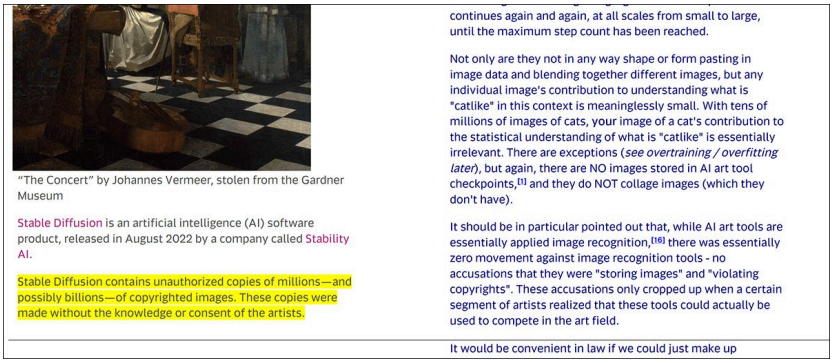
Further duplicate reproductions that proved possible during the researchers' experiments. See original paper for better resolution, but the 'copes' are remarkably faithful.

They note also prior work indicating that *memorization*, typically a pejorative term, may actually be *necessary* to enable the ground-breaking performance of latent diffusion-based generative models. This raises the question, the researchers state, as to 'whether the improved performance of diffusion models compared to prior approaches is precisely *because* diffusion models memorize more'.

Copycat AI?

If the paper's findings in this respect are borne out over time, they could notably influence current lawsuits related to the alleged plagiaristic capacities of Stable Diffusion. On 13th of January this year, noted AI commenter Matthew Butterick announced the launch of a class-action complaint against Stability.ai, the creators of Stable Diffusion, in the context of a dedicated site arguing that Stable Diffusion is essentially a 'copy and paste' framework that infringes on the work of original creators.

Some days later a rebuttal domain was launched, arguing instead that the litigants had failed to understand how generalization works, and that Stable Diffusion (and similar systems), actually functions by observing and learning principles from data, similar to the way an artist or writer may appreciate many paintings or read many books before going on to produce their own original works, where the 'seen' data operates as an 'influence', rather than being substantially reproduced.



From the original Butterick domain (https://stablediffusionlitigation.com/, left) and the answer domain (http://www.stablediffusionfrivolous.com/), two differing takes on whether Stable Diffusion is ripping artists off or being inspired by them.

The new paper states:

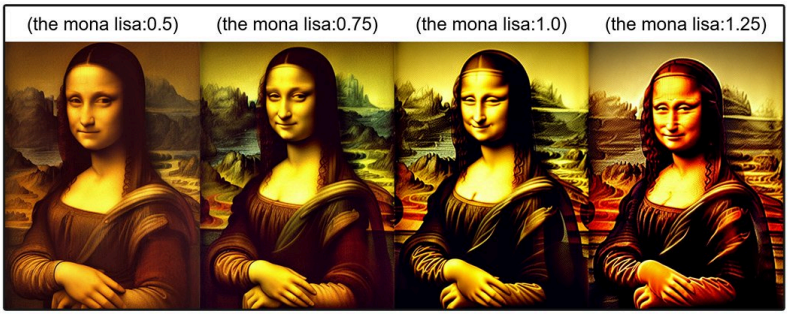
'Do large-scale models work by generating novel output, or do they just copy and interpolate between individual training examples? If our extraction attacks had failed, it may have refuted the hypothesis that models copy and interpolate training data; but because our attacks succeed, this question remains open.

'Given that different models memorize varying amounts of data, we hope future work will explore how diffusion models copy from their training datasets.'

Duplication and Outliers

LAION-scale datasets are minimally curated, at least through actual human oversight, and the challenge of getting intelligible results and coherent synthesis from the text/image pairs that constitute the data is in managing and mitigating the quality of labeling and the inevitable sheer number of duplicate images that are likely to exist in a dataset scraped from the web at hyperscale, leading to an otherwise unmanageable collection of billions of images.

In the case of popular images, such as the Mona Lisa, this means that certain images are likely to appear in the under-curated dataset multiple times. The more often an AI model is exposed to an essentially identical image as it iterates through the dataset during training, the more likely it is to memorize the image, and to learn to reproduce it perfectly if an apposite prompt is given at inference time.



Though a default prompt of 'Mona Lisa' will elicit a low-quality image in Stable Diffusion, changing the CFG parameters can obtain a more faithful reproduction. Source: https://old.reddit.com/r/StableDiffusion/comments/zgm7a4/no_matter_which_version_of_stable_diffusion_or/

By this same process, Stable Diffusion will learn best the features that recur most frequently, which is why it is able to quite accurately reproduce the most popular celebrities of the last 5-10 years (or those older celebrities that remain iconic, and whose images still circulate widely in the culture) without the use of DreamBooth or other fine-tuning techniques. However, in this case, the system's native ability to 'deepfake' a celebrity is more generalized; due to the sheer volume and variety of source images in the LAION subset, no particular image is likely to predominate or to appear *verbatim*, as it were.

To gain better insight into the process by which source data becomes reproducible, the authors custom-trained a number of models on the CIFAR-10 dataset, which brought out additional insights:

'We also surprisingly find that diffusion models and GANs memorize many of the same images. In particular, despite the fact that our diffusion model memorizes 1280 images and a StyleGAN model we train on half of the dataset memorizes 361 images, we find that 244 unique images are memorized in common.

'If images were memorized uniformly at random, we should expect on average 10 images would be memorized by both, giving exceptionally strong evidence that some [images] are inherently less private than others. Understanding why this phenomenon occurs is a fruitful direction for future work.'

The researchers also found that painstaking deduplication did not mitigate the problem as much as might be expected. Having removed duplicates from the CIFAR-10 dataset on which they were training, the incidence of memorized images extractable from the resulting diffusion model fell from 1280 to 986, whereas logic suggested the improvement would be far higher, and the paper comments:

'While not a substantial drop, these results show that deduplication can mitigate memorization. Moreover, we also expect that deduplication will be much more effective for models trained on larger-scale datasets (e.g., Stable Diffusion), as we observed a much stronger correlation between data extraction and duplication rates for those models.'

Therefore the researchers adjure the sector to commit to more sophisticated deduplication techniques, rather than simply de-duping by not taking the same image twice from the same URL, and to evaluate similarity via CLIP and similar metrics and approaches.

Imagen the Greater 'Thief'

Since few images reach the distribution level of the Mona Lisa, it could be expected that generalization itself will protect a latent diffusion system from becoming capable of spitting out accurate reproductions of source data. Aware of this, the researchers investigated the extent to which outlier data (i.e., data that is fairly specific and unusual) can be re-extracted from a diffusion model.

The hard way to prove this would have been to essentially train a novel Stable Diffusion-scale model from scratch, and follow the progress of individual images. Since this is not a practical (or at least casually reproducible) solution, the researchers instead computed the CLIP embedding of training examples, seeking what they call the 'outlierness' quality. They found, unexpectedly, that Google's Imagen synthesis framework produced very different statistics to Stable Diffusion in this regard*:

'Surprisingly, we find that attacking out-of-distribution images is much more effective for Imagen than it is for Stable Diffusion. On Imagen, we attempted extraction of the 500 images with the highest out-of-distribution score. Imagen memorized and regurgitated 3 of these images (which were unique in the training dataset).

'In contrast, we failed to identify any memorization when applying the same methodology to Stable Diffusion – even after attempting to extract the 10,000 most-outlier samples. Thus, Imagen appears less private than Stable Diffusion both on duplicated and non-duplicated images.

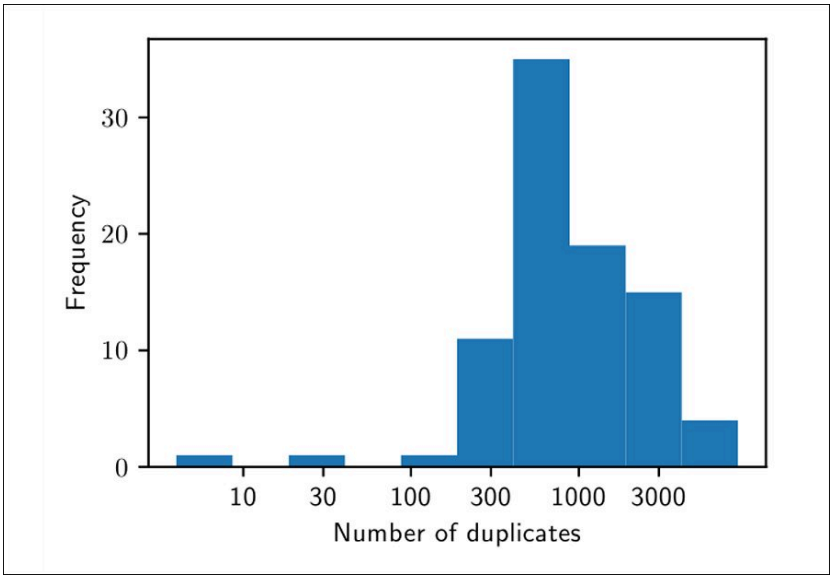
'We believe this is due to the fact that Imagen uses a model with a much higher capacity compared to Stable diffusion, which allows for more memorization. Moreover, Imagen is trained for more iterations and on a smaller dataset, which can also result in higher memorization.'

The majority of images reproduced from training data in the researchers' experiments were photographs of a recognizable person, the remainder being commercial products (17%), logos and posters (14%) and other art or graphics; and its notable that the remaining types of subject matter seem likely to be relatively invariant across examples, and to be duplicated, if only due to the excess of advertising, product placement and astroturfing to be found in a typical web-scrape.

Before conducting the tests, the researchers made some attempt to assess the extent of duplication in the data, accounting for factors such as compression quality and the inclusion of an image in a larger image, and seeking instead to identify 'near duplication'.

Having identified the 350,000 most-duplicated examples from the source dataset, the researchers generated a staggering 500 candidate images for each related text-prompt, to arrive at a total of 175 million generated images.

The images were then ordered by their mean distance to each other, since the closer the highest number of similar instances are, the more likely they are to become memorized and accessible in a deployed system.



The paper notes that the 'attacks' (extraction methods) used in the study produces the most number of successes when the source image has been duplicated at least 100 times in the dataset, though the authors caution that this is an upper bound and not the lower bound of a higher optimum number, since they were specifically seeking duplicated images.

The authors assert:

'Our attack is exceptionally precise: out of 175 million generated images, we can identify 50 memorized images with 0 false positives, and all our memorized images can be extracted with a precision above 50%.'

Reproductions

Though the paper's findings may be of most interest to current and future litigants of Stable Diffusion (Imagen has never been made available, either as an open source architecture or via an API), the paper emphasizes that the results have negative indications also for the possible use of diffusion models in systems that process personally identifiable information (PII), or where the source training content may be in some or other way sensitive, and where attackers should not have an easy way to extract specific data from the final model.

Though there are a variety of techniques that can help to preserve sensitive information (and even provenance of information) during training, the researchers found that not all are usefully applicable. For instance, the use of differentially-private stochastic gradient descent (DP-SGD), when tried by the researchers in the training of a test diffusion model, caused the CIFAR-10-based model to 'consistently diverge' from optimal goals.

'In fact,' the researchers note. 'even applying a non-trivial gradient clipping or noising on their own (both are required in DP-SGD) caused the training to fail. We leave a further investigation of these failures to future work, and we believe that new advances in DP-SGD and privacy-preserving training techniques may be required to train diffusion models in privacy-sensitive settings.'