Joshua Misir, Patrick Mazza
CS 6382 - Data Science for Social Change
Professor Pierson
Final Report

Machine learning models are increasingly being used in a wide range of applications, including image and speech recognition, natural language processing, and autonomous driving. One type of machine learning that is becoming increasingly popular is generative machine learning. Generative models are able to create new data that is similar to the data that they were trained on. This makes them useful for a variety of tasks, such as generating new images, text, and music. One type of generative model that is gaining popularity is the diffusion model. Diffusion models work by gradually adding noise to an image or text corpus. The model then learns to remove the noise and recover the original image or text. Diffusion models have been shown to be able to generate high-quality images and text, and they are relatively easy to train. However, diffusion models also have some security risks. One risk is that they can be used to extract training data from the model. This can be done by prompting the model to generate images or text that are similar to images or text that are present in the training data. If an adversary is able to extract training data from a diffusion model, they could potentially use that data to identify individuals or find out sensitive information about them that would otherwise be known. This article will discuss the security risks of diffusion models and how to mitigate those risks. We will also discuss the implications of these risks for the future of diffusion models.

To get a better understanding of the process that operates under the hood of a diffusion model, let's say you want to train a machine learning model to create images of birds. You have a clean image of a bird and you then add some type of noise to the image such as visual distortions. Next, you ask the machine learning model to denoise the model by removing any distortions, resulting in a clean image of the bird. By denoising many images, with various types and extents of noise, the machine learning model gets better at cleaning up noisy images and it learns about the patterns and features that make up realistic images. The model's insight on these patterns and features allows it to generate new, high-quality images that may appear indistinguishable from a real photo. As the machine learning model becomes proficient at denoising images by learning patterns and features that constitute real images, it gains the ability to generate new, high quality images by extrapolating from the learned knowledge of image structures and characteristics.

Machine learning models can inadvertently leak details of their training data, posing risks to individuals' privacy and introducing other concerns. Two specific types of attacks, inversion attacks and attribute attacks, highlight these vulnerabilities. It's important to note that while some attacks may not directly relate to privacy, they still raise concerns about data misuse.

Inversion attacks focus on the recovery of training data used to train a model. For instance, in a picture generation model, if a user can partially or fully reconstruct the original training examples, it becomes a privacy threat, especially if sensitive information is present in the training data. This type of attack allows adversaries to gain access to potentially private details through the model's outputs.

Attribute attacks, on the other hand, exploit biases and unintended correlations learned by machine learning models. For example, let's consider a financial loan model that predicts loan eligibility. An adversary with access to basic information about an individual, such as their age and occupation, can manipulate the data by changing only the gender attribute. If the model consistently approves loans based on male gender, the adversary can infer the gender of targeted individuals. This attack demonstrates how models unintentionally learn and exploit sensitive attributes, even if those attributes were not explicitly included in the prediction task. Copyright issues can also arise when diffusion models generate public data that includes copyrighted text, images, or source code. These challenges raise legal concerns around the usage and ownership of intellectual property.

Training data extraction is said to be memorized if a sequence of words in a text model can be recovered, *verbatim*, by an adversary prompting the model. The same is true for image models where an adversary can prompt a model and receive an exact or near identical image which is present in the data set. The main concern isn’t the model generating similar types of images based on what it was trained on. Instead, the main concern that poses security risks is the fact that the generative model may output a near identical image that was present in the training data set, especially when images that aren't meant to be publicly available find themselves in the training data. In theory, all non-public images that may contain sensitive information shouldnt find be on the public internet, exposed to the potential to be included in a machine learning models training set. In practice, this is not the case. LAION (Large-Scale Artificial Intelligence Open Network), for example, comprises various types of sensitive information such as patients' medical images.

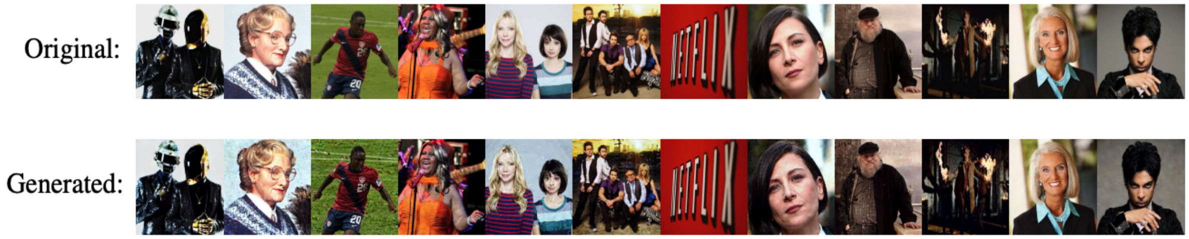

Figure 3: Examples of the images that we extract from Stable Diffusion v1.4 using random sampling and our membership inference procedure. The top row shows the original images and the bottom row shows our extracted images.

Recent research exploring the security concerns related to image diffusion models was released in a paper titled "Extracting Training Data from Diffusion Models", published in January 2023, by researchers from Google, DeepMind, ETHZ, Princeton and UC Berkeley. Researchers took the following steps to identify how much memorization is present in diffusion models:

1. Using Stable Diffusion, generate many images examples via prompts
2. Perform membership inference by determining which generated images are novel and which appear to be memorized from the training set. (e.g. if 10 or more images are generated from a single prompt that are near identical, its predicted that the images are memorized
3. After 175 million image generations based on the most duplicated examples, these generated images are then compared to the images present in the data set by computing Euclidean distance of the images and comparing a small subset of the images manually.

The results of this process is that a total of 109 image generations are near-identical training set images. A key finding also highlights that the higher number of duplicates that are present in the training dataset, the more likely it is to be memorized and potentially subject to an attack.

The same procedure was done using Imagen, another popular text-to-image generation model. The results were even more profound when attacking the Imagen model, which indicates it may be less private than Stable Diffusion. 23 images from 500,000 total images were identified to be memorized from the training dataset. This may sound like a small number of memorized images, given the size of the total amount of images. This is due this being the first study exploring the possibility of such vulnerabilities in diffusion models using a strict definition of memorization (near identical image reconstruction). Additionally, these results demonstrate that memorization across diffusion models is highly dependent on training configurations such as model size, dataset size, and training time. One potential explanation for why Imagen may be less private than Stable Diffusion is due to the fact that Imagen is trained on more iterations of smaller dataset, which can lead to an increase in the tendency for a machine learning model to merely memorize training examples. The majority of photographs analyzed feature a recognizable person as the primary subject, while others depict products, logos, or art. If a future diffusion model were trained on sensitive data, the extracted information would likely come from that sensitive (e.g. medical, financial) data distribution. Although these images are publicly accessible online, not all are permissively licensed. Of the results, 35% fall under non-permissive copyright notice and 61% potentially under general copyright protection. A few images are licensed under CC BY-SA, requiring proper credit, a link to the license, and indication of any changes made.

This work done in the paper "Extracting Training Data from Diffusion Models" sheds light on serious security concern for diffusion models. The authors demonstrate that it is possible to extract training data from diffusion models, even when the models are trained on large datasets. This could allow adversaries to train rival models that are able to generate the same types of images or text. The authors also note that the risk of these attacks increases as the number of duplicates in the training dataset increases. This research suggests that diffusion models are not as secure as previously thought, and that they may be vulnerable to attacks that aim to extract training data. This is a serious concern, as it could have implications for the future of diffusion models. It is important to be aware of this security concern and to take steps to mitigate it, such as using privacy-preserving training techniques.