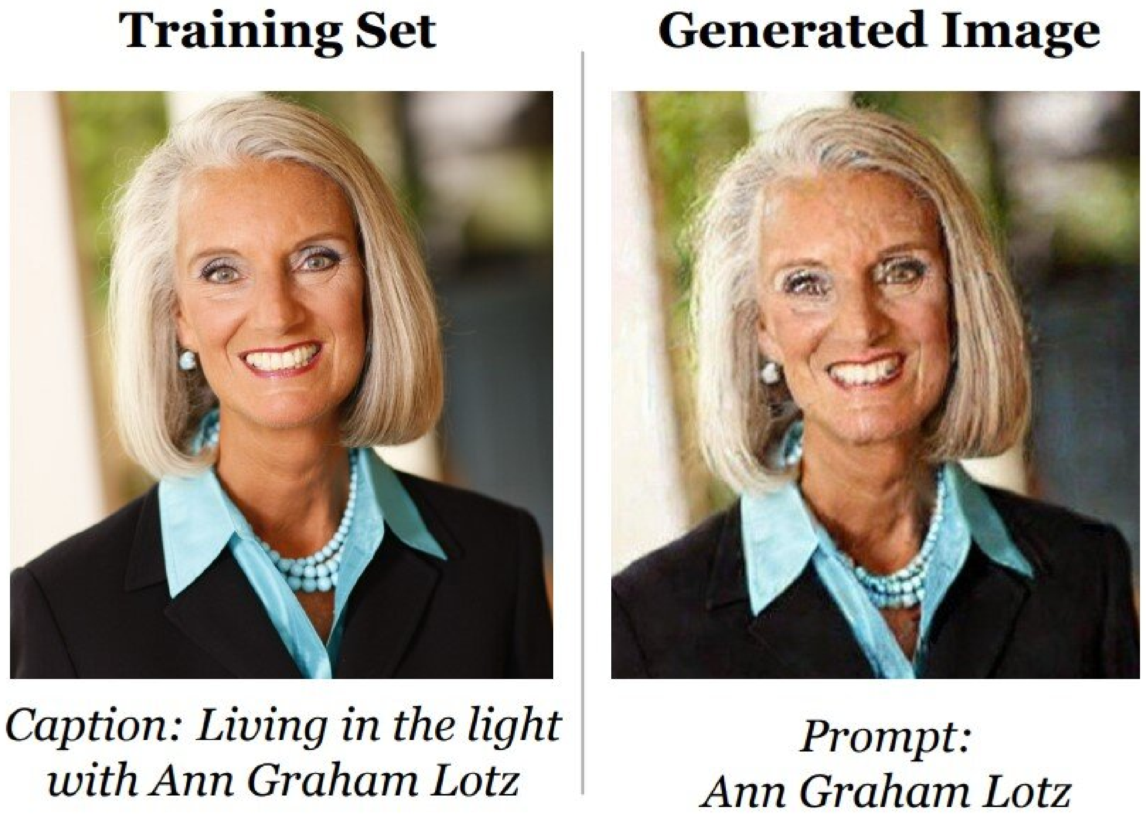
TechXplore

Testing shows AI-based image generation systems can sometimes generate copies of trainer data

February 2 2023, by Bob Yirka



Diffusion models memorize individual training examples and generate them at test time. Left: an image from Stable Diffusion’s training set (licensed CC BY-SA 3.0). Right: a Stable Diffusion generation when prompted with “Ann Graham Lotz”. The reconstruction is nearly identical (`2 distance = 0.031). Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2301.13188

A team of computer scientists from Google, DeepMind, ETHZ, Princeton University and the University of California, Berkeley, has found that AI-based image generation systems can sometimes generate copies of images used to train them. The group has published a paper describing testing of several image generation software systems on the *arXiv* preprint server.

Image generation systems such as Stable Diffusion, Imagen and Dall-E 2 have been in the news lately due to their ability to generate high-resolution images based on nothing but natural language prompts. Such systems have been trained on thousands of images as templates.

In this new effort, the researchers, some of whom are part of a team that created one of the systems, have found that these systems can sometimes make a pretty important mistake. Instead of generating a new image, the system simply spits out one of the images from its training data. It happens somewhat frequently—they found more than 100 instances out of 1,000 image returns during their testing efforts.

This is a problem because the datasets are typically scraped from the internet, and many carry copyrights. During testing, the team found that approximately 35% of the copied images had copyright notices. Approximately 65% did not have an explicit notice, but appeared likely to belong to images covered under general copyright protection laws.

The researchers note that most AI-based image generation systems have a processing stage during which noise is added to prevent the return of images from datasets, pushing the system to create something new. They also note that sometimes a system added noise to a copied image, making it more difficult to tell that it was a copy.

The team concludes that producers of such products need to add another safeguard to prevent copies from being returned. They note a simple flagging mechanism should do the trick.

More information: Nicholas Carlini et al, Extracting Training Data from Diffusion Models, *arXiv* (2023). DOI: 10.48550/arxiv.2301.13188

Citation: Testing shows AI-based image generation systems can sometimes generate copies of trainer data (2023, February 2) retrieved 20 March 2024 from https://techxplore.com/news/2023-02-ai-based-image-generation-generate-trainer.html