A Meta-Study on Methods of Poisoning Artificial Intelligence Text-to-Image Generators

CS4991 Capstone Report, 2024

Nick Garrone
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
nkj9hk@virginia.edu

ABSTRACT

Some artists wish to prevent their work from being used to train AI-powered text-to-image generators such as Stable Diffusion in order to protect the integrity of their work. Artists may poison their images using a tool that modifies the image in such a way that it can no longer be used as reliable training data. This meta-study review examines several methods of poisoning images, including bad data labeling, The University of Chicago's Glaze and Nightshade systems, and backdoor injection, in order to determine how these tools can best serve the needs of artists. I found that methods of poisoning can be prompt-specific or general, require large or small datasets, and have an impact on the image or have no impact. Each has use in specific situations– some methods are able to be used by a single artist to affect another's model while others would require the distribution of poisoned models. As text-to-image models may be trained to prevent attacks, these methods must be continuously iterated upon to stay effective.

1. INTRODUCTION

Within the past three years, artificial intelligence text-to-image generators have rapidly advanced. These tools allow users to generate images in a wide variety of styles, including near-photorealism, with just a short text prompt. In 2021, OpenAI released the first text-to-image model Dalle-E. Its successor, Dalle-2, was the first diffusion-based model. An open-source tool based on this architecture, Stable Diffusion, has also seen popularity [1].

The growth of these tools have sparked concern among artists. Not only do they have the possibility of supplanting artists in the workplace, but some models are trained on copyrighted works. An unresolved lawsuit, *Andersen v. Stability AI et al.* alleges that the training of AI on their content constitutes a violation of copyright [2]. Regardless of the outcome of the legal challenges, many artists want to actively prevent their work from being used to train these models. A variety of methods can serve that need.

2. BACKGROUND

The current state of the art in image generation is based on Latent Diffusion Models. Rombach, et al. (2022) propose this method. Diffusion Models work by applying denoising autoencoders, machine learning models designed to compress images by reducing the noise in an image, sequentially. By gradually denoising a random image according to some prompt, the diffusion model actually creates a new image. The "latent" part of the model is a breakthrough that has rapidly improved the quality of images. Latent Diffusion Models are models in which the denoising occurs in latent, or compressed, space before being translated back to pixel space. The compression is done by another autoencoder [3].

3. REVIEW OF RESEARCH

A naïve method of poisoning, "dirty-labeling," is described by Shan, et al. This entails incorrectly labeling the contents of images, en masse [4]. A more sophisticated method is proposed in the same paper. The goal of Nightshade is to produce similar outcomes to the dirty-label attack, but to do so with less data and to hide the corruption of the images [4].

Whereas Nightshade aims to produce incorrect output for a given concept, Struppek, et al. describe a method for generating unexpected images based on a specific textual character rather than a concept. This backdoor can take the form of an uncommon or non-latin character. This could be used to generate unexpected images based on an uncommonly used character. This method requires modifying the model itself [5].

Another system, Glaze, seeks to protect certain styles rather than concepts. Similar to Nightshade, it applies barely-visible alterations to the image. However, unlike Nightshade, it is designed to protect style mimicry in particular, and it targets a different stage of the training pipeline [6].

4. META-STUDY ANALYSIS

These methods of poisoning models function by targeting the training step of the pipeline. Because the goal is to prevent their images from being used in the training step of the pipeline, these methods all involve corrupting the data in some way. The main difference between the methods is the sophistication with which they corrupt the data and in what conditions the effects of the corruption are seen. The simplest method of corrupting the data is by dirty-labeling it. This entails labeling an image with an incorrect label. Because the training process uses the label to determine which concept the image represents, an incorrect label would cause the model to have an incorrect understanding of the image. The simplicity comes with a cost—human inspection or an image recognition model could easily see that the image does not match its label [4].

Nightshade proposes a much more sophisticated method of corrupting the data. In order to avoid detection, Nightshade solves an optimization problem in order to minimize the perturbations to a target image while maximizing the potency of the poison. The new image it creates through this process is nearly indistinguishable from the original [4].

To do this, Nightshade takes advantage of a fundamental inefficiency of model training. Consider the concept "dog." There are countless ways that this concept can be represented in training data: a picture of a golden retriever running through a field, a cartoon image of Scooby-Doo, a painting of a labradoodle. The way each image contributes to the model's understanding of "dog" is noisy and often contradictory. The images can be said to be unaligned. To take advantage of this unalignment, Nightshade ensures that its poisoned images are as aligned as possible. Each poison image tries to trick the model in the same way. Additionally, each poison image is originally generated by querying the target model for the fake concept, meaning that it is guaranteed to be aligned with the model's ground understanding of the concept. This allows for the poison images to be more potent than the benign images [4].

Glaze takes a similar approach. Instead of targeting a concept, it targets a style. It aims to improve the efficacy of protecting styles by optimizing for protecting against style mimicry over protecting other parts of an image. To do this, a target image is first transformed by the target model to have a different style. For example, a photorealistic image becomes like a painting. Then, it uses

the transformed image in order to identify the style-specific features in the original image, which is then focused on as the image is perturbed. Like Nightshade, the perturbation aims to optimize potency while minimizing visual change [6].

A final technique is backdoor injection. Rather than targeting a concept or style, it targets a specific character, often an odd or rarely-used character. In this way it can produce unexpected images, reducing trust in the models. This method, rather than modifying images, requires the modification of the encoder of the model.

It does this using a teacher-student architecture: two initially identical encoders are created, a teacher and a student. The student encoder is changed to inject the backdoor into the embedding, or numerical representation of the image, given the non-latin character's presence in a prompt, while the teacher model verifies that the student produces correct output for normal prompts. This functions more as a proof-of-concept than anything else: it requires modification to the model itself, which cannot be done via artists when treating their images. To release this attack into the wild would require distributing the poisoned model to end users rather than modifying another person's model [5].

5. META-STUDY FINDINGS

In fulfilling the goal of protecting artists' work from being used in training, Nightshade provides the most effective and robust protection. For artists who are specifically looking to protect their style, Glaze is a better choice. Dirty-labeling data, while simple, is not as effective. The use of character-based backdoors, while it succeeds in the goal of modifying the model in undesired ways, does not truly provide robust protection. Additionally, it requires access to the model itself, which severely limits the attack vectors for artists.

6. CONCLUSION

With more knowledge of the methods available to them, artists will be able to more effectively protect their artwork as text-to-image models proliferate. Wider use of these tools would impede the development of text-to-image models trained on the open internet. Additionally, this research provides a starting point for computer scientists to learn about the fundamentals of image poisoning.

7. FUTURE WORK

To deepen understanding of the methods presented, future work could actively test the methods presented here by fine-tuning a pre-trained model on images poisoned by different methods. This would provide a more quantitative look into the efficacy of the methods. Additionally, as text-to-image models advance, they may be able to counteract the methods of poisoning described in this report. In order to remain relevant, future work may examine new methods as they are developed, comparing them to the methods presented here.

REFERENCES

[1] Tom Faber. 2022. The golden age of AI-generated art is here. It's going to get weird. Retrieved April 12, 2024 from https://www.ft.com/content/073ea888-20d7-437c-8226-a2dd9f276de4

[2] Blake Brittain. 2023. Artists take new shot at Stability, Midjourney in updated copyright lawsuit. Retrieved April 12, 2024 from https://www.reuters.com/legal/litigation/artists-take-new-shot-stability-midjourney-updated-copyright-lawsuit-2023-11-30/

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2021. High-Resolution Image

Synthesis with Latent Diffusion Models. arXiv:2112.10752. Retrieved from https://arxiv.org/pdf/2112.10752.pdf

[4] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. 2024. Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. arXiv:2310.13828. Retrieved from https://arxiv.org/pdf/2310.13828.pdf

[5] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2023. Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. October 2-3, 2023, Paris, France. https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00423

[6] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. In *Proceedings of the 32nd USENIX Security Symposium*. August 9-1, 2023, Anaheim, CA. https://www.usenix.org/system/files/usenixsecurity23-shan.pdf