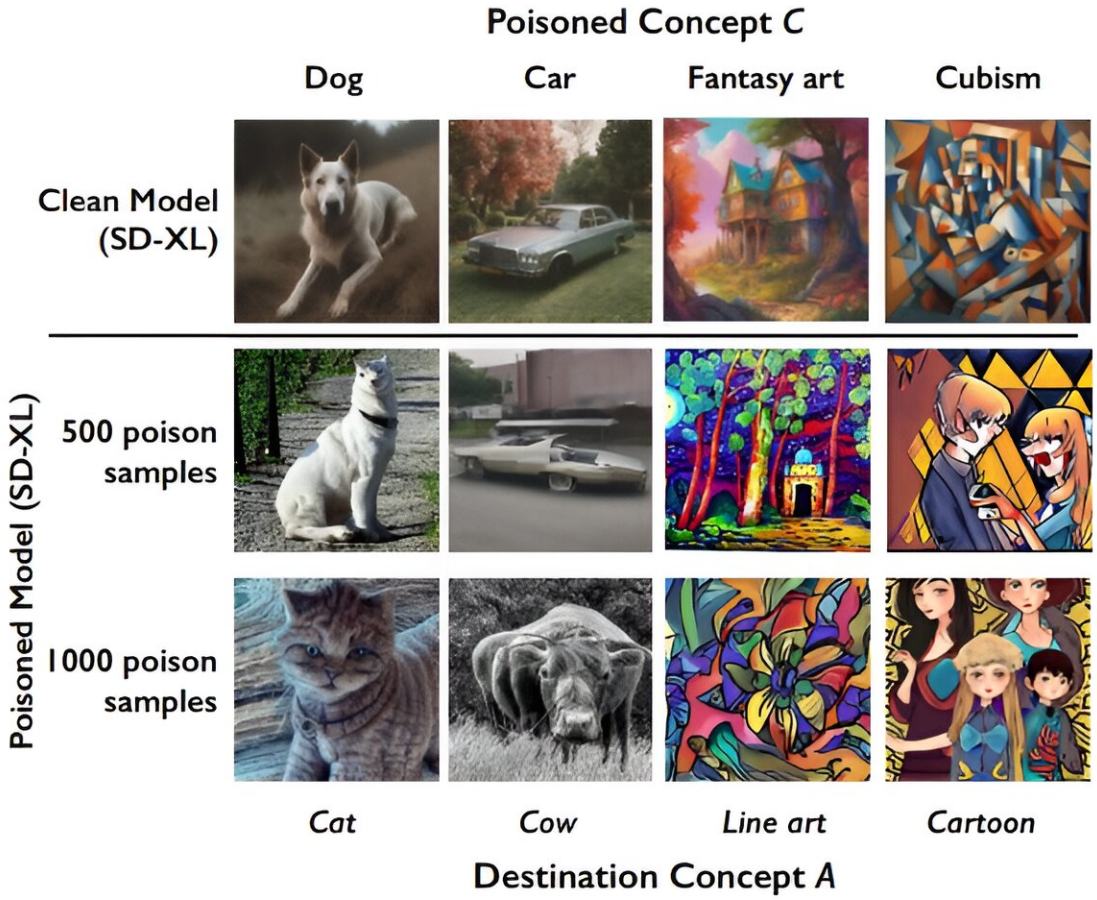
Sabotage tool takes on AI image scrapers

October 25 2023, by Peter Grad



Example images generated by the clean (unpoisoned) and poisoned SD-XL models with different # of poison data. The attack effect is apparent with 1000 poisoning samples, but not at 500 samples. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.13828

Artists who have stood by helplessly as their online works remained ripe for the picking without authorization by AI web scraping operations can finally fight back.

Researchers at the University of Chicago announced the development of a tool that "poisons" graphics appropriated by AI companies to train image-generating models. The tool, Nightshade, manipulates image pixels that will alter the output during training. The alterations are not visible to the naked eye prior to processing.

Ben Zhao, an author of the paper "Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models," said Nightshade can sabotage data so that images of dogs, for instance, would be converted to cats at training time. In other instances, car images were transformed into cars, and hats converted to cakes. The work is published on the *arXiv* preprint server.

"A moderate number of Nightshade attacks can destabilize general features in a text-to-image generative model, effectively disabling its ability to generate meaningful images," Zhao said.

He termed his team's creation "a last defense for content creators against web scrapers that ignore opt-out/do-not-crawl directives."

Artists have long worried about companies such as Google, OpenAI, Stability AI and Meta that collect billions of images online for use in training datasets for lucrative image-generating tools while failing to provide compensation to creators.

Eva Toorenent, an adviser for the European Guild for Artificial Intelligence Regulation in the Netherlands, said such practices "have sucked the creative juices of millions of artists."

"It is absolutely horrifying," she said in a recent interview.

Zhao's team demonstrated that despite the common belief that disrupting scraping operations would require uploading massive amounts of altered images, they were able to achieve disruption by using fewer than 100 "poisoned" samples. They achieved this by using prompt-specific poisoning attacks that require far fewer samples than the model training dataset.

Zhao sees Nightshade as a useful tool not only for individual artists but for large companies as well, such as movie studios and game developers.

"For example, Disney might apply Nightshade to its print images of 'Cinderella,' while coordinating with others on poison concepts for 'Mermaid,'" Zhao said.

Nightshade can also alter art styles. For instance, a prompt requesting an image be created in Baroque style may yield Cubist style imagery instead.

The tool emerges in the midst of rising opposition to AI companies appropriating web content under what the companies say is allowed by fair-use rules. Lawsuits were filed against Google and Microsoft's OpenAI last summer accusing the tech giants of improperly using copyrighted materials to train their AI systems.

"Google does not own the internet, it does not own our creative works, it does not own our expressions of our personhood, pictures of our families and children, or anything else simply because we share it online," said the plaintiffs' attorney, Ryan Clarkson. If found guilty, the companies face billions in fines.

Google seeks a dismissal of the lawsuit, stating in court papers, "Using

publicly available information to learn is not stealing, nor is it an invasion of privacy, conversion, negligence, unfair competition, or copyright infringement."

According to Toorenent, Nightshade "is going to make [AI companies] think twice, because they have the possibility of destroying their entire model by taking our work without our consent."

More information: Shawn Shan et al, Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models, *arXiv* (2023). DOI: 10.48550/arxiv.2310.13828