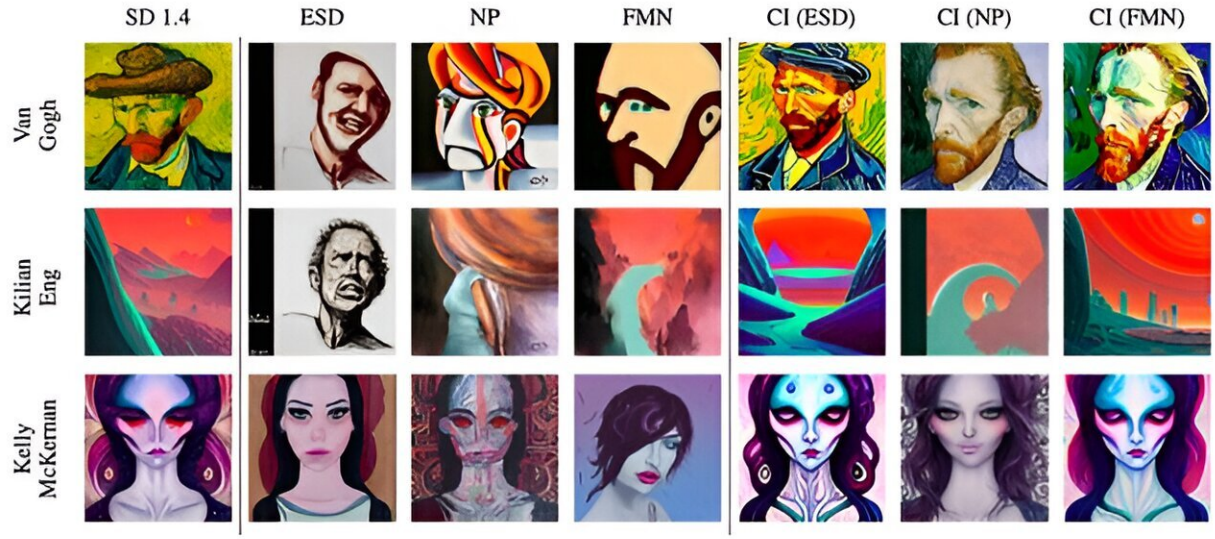
Study exposes failings of measures to prevent illegal content generation by text-to-image AI models

March 14 2024



Concept Inversion (CI) on Erased Stable Diffusion (ESD), Negative Prompt (NP) and Forget-Me-Not (FMN) for art concept. The first three columns demonstrate the effectiveness of concept erasure methods when using the prompt: "a painting in the style of [artist name]". However, when the researchers replace [artist name] with the special token learned by Concept Inversion, the model can still generate images of the erased styles. Credit: Minh Pham et al

Researchers at NYU Tandon School of Engineering have revealed critical shortcomings in recently-proposed methods aimed at making powerful text-to-image generative AI systems safer for public use.

In a paper that will be presented at the Twelfth International Conference on Learning Representations (ICLR), taking place in Vienna on May 7–11, 2024, the research team demonstrates how techniques that claim to "erase" the ability of models like Stable Diffusion to generate explicit, copyrighted, or otherwise unsafe visual content can be circumvented through simple attacks. The paper also appears on the pre-print server *arXiv*.

Stable Diffusion is a publicly available AI system that can create highly realistic images from just text descriptions. Examples of the images generated in the study are on GitHub.

"Text-to-image models have taken the world by storm with their ability to create virtually any visual scene from just textual descriptions," said the paper's senior author Chinmay Hegde, associate professor in the NYU Tandon Electrical and Computer Engineering Department and in the Computer Science and Engineering Department. "But that opens the door to people making and distributing photo-realistic images that may be deeply manipulative, offensive and even illegal, including celebrity deepfakes or images that violate copyrights."

The researchers investigated seven of the latest concept erasure methods and demonstrated how they could bypass the filters using "concept inversion" attacks.

By learning special word embeddings and providing them as inputs, the researchers could successfully trigger Stable Diffusion to reconstruct the very concepts the sanitization aimed to remove, including hate symbols, trademarked objects, or celebrity likenesses. In fact the team's inversion

attacks could reconstruct virtually any unsafe imagery the original Stable Diffusion model was capable of, despite claims the concepts were "erased."

The methods appear to be performing simple input filtering rather than truly removing unsafe knowledge representations. An adversary could potentially use these same concept inversion prompts on publicly released sanitized models to generate harmful or illegal content.

The findings raise concerns about prematurely deploying these sanitization approaches as a safety solution for powerful generative AI.

"Rendering text-to-image generative AI models incapable of creating bad content requires altering the model training itself, rather than relying on post hoc fixes," said Hegde. "Our work shows that it is very unlikely that, say, Brad Pitt could ever successfully request that his appearance be 'forgotten' by modern AI. Once these AI models reliably learn concepts, it is virtually impossible to fully excise any one concept from them."

According to Hegde, the research also shows that proposed concept erasure methods must be evaluated not just on general samples, but explicitly against adversarial concept inversion attacks during the assessment process.

Collaborating with Hegde on the study were the paper's first author, NYU Tandon Ph.D. candidate Minh Pham; NYU Tandon Ph.D. candidate Govin Mittal; NYU Tandon graduate fellow Kelly O. Marshall and NYU Tandon post doctoral researcher Niv Cohen.

The paper is the latest research that contributes to Hegde's body of work focused on developing AI models to solve problems in areas like imaging, materials design, and transportation, and on identifying weaknesses in current models.

In another recent study, Hegde and his collaborators revealed they developed an AI technique that can change a person's apparent age in images while maintaining their unique identifying features, a significant step forward from standard AI models that can make people look younger or older but fail to retain their individual biometric identifiers.

More information: Minh Pham et al, Circumventing Concept Erasure Methods For Text-to-Image Generative Models, *arXiv* (2023). DOI: 10.48550/arxiv.2308.01508

Provided by NYU Tandon School of Engineering