Using AI to protect against AI image manipulation

July 31 2023, by Rachel Gordon

In this example, an adversary seeks to modify an image found online. The adversary describes via textual prompt the desired changes and then uses a diffusion model to generate a realistic image that matches the prompt. By immunizing the original image before an adversary can access it, the PhotoGuard system disrupts the ability to successfully perform such edits. Credit: Massachusetts Institute of Technology

As we enter a new era where technologies powered by artificial intelligence can craft and manipulate images with a precision that blurs the line between reality and fabrication, the specter of misuse looms large.

Recently, advanced generative models such as DALL-E and Midjourney, celebrated for their impressive precision and user-friendly interfaces, have made the production of hyper-realistic images relatively effortless. With the barriers of entry lowered, even inexperienced users can generate and manipulate high-quality images from simple text descriptions—ranging from innocent image alterations to malicious changes.

Techniques like watermarking pose a promising solution, but misuse requires a preemptive (as opposed to only post hoc) measure.

In the quest to create such a new measure, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) developed "PhotoGuard," a technique that uses perturbations—minuscule alterations in pixel values invisible to the human eye but detectable by computer models—that effectively disrupt the model's ability to manipulate the image.

PhotoGuard uses two different "attack" methods to generate these perturbations. The more straightforward "encoder" attack targets the image's latent representation in the AI model, causing the model to perceive the image as a random entity. The more sophisticated "diffusion" one defines a target image and optimizes the perturbations to make the final image resemble the target as closely as possible.

"Consider the possibility of fraudulent propagation of fake catastrophic events, like an explosion at a significant landmark. This deception can manipulate market trends and public sentiment, but the risks are not

limited to the public sphere. Personal images can be inappropriately altered and used for blackmail, resulting in significant financial implications when executed on a large scale," says Hadi Salman, an MIT graduate student in electrical engineering and computer science (EECS), affiliate of MIT CSAIL, and lead author of a new paper about PhotoGuard available on the *arXiv* preprint server.

"In more extreme scenarios, these models could simulate voices and images for staging false crimes, inflicting psychological distress and financial loss. The swift nature of these actions compounds the problem. Even when the deception is eventually uncovered, the damage—whether reputational, emotional, or financial—has often already happened. This is a reality for victims at all levels, from individuals bullied at school to society-wide manipulation."

PhotoGuard in practice

AI models view an image differently from the way humans do. It sees an image as a complex set of mathematical data points that describe every pixel's color and position—this is the image's latent representation. The encoder attack introduces minor adjustments into this mathematical representation, causing the AI model to perceive the image as a random entity.

As a result, any attempt to manipulate the image using the model becomes nearly impossible. The changes introduced are so minute that they are invisible to the human eye, thus preserving the image's visual integrity while ensuring its protection.

The second and decidedly more intricate "diffusion" attack strategically targets the entire diffusion model end-to-end. This involves determining a desired target image, and then initiating an optimization process with the intention of closely aligning the generated image with this

preselected target.

In implementing, the team created perturbations within the input space of the original image. These perturbations are then used during the inference stage, and applied to the images, offering a robust defense against unauthorized manipulation.

"The progress in AI that we are witnessing is truly breathtaking, but it enables beneficial and malicious uses of AI alike," says MIT professor of EECS and CSAIL principal investigator Aleksander Madry, who is also an author on the paper. "It is thus urgent that we work towards identifying and mitigating the latter. I view PhotoGuard as our small contribution to that important effort."

The diffusion attack is more computationally intensive than its simpler sibling, and requires significant GPU memory. The team says that approximating the diffusion process with fewer steps mitigates the issue, thus making the technique more practical.

To better illustrate the attack, consider an art project, for example. The original image is a drawing, and the target image is another drawing that's completely different. The diffusion attack is like making tiny, invisible changes to the first drawing so that, to an AI model, it begins to resemble the second drawing. However, to the human eye, the original drawing remains unchanged.

By doing this, any AI model attempting to modify the original image will now inadvertently make changes as if dealing with the target image, thereby protecting the original image from intended manipulation. The result is a picture that remains visually unaltered for human observers, but protects against unauthorized edits by AI models.

As far as a real example with PhotoGuard, consider an image with

multiple faces. You could mask any faces you don't want to modify, and then prompt with "two men attending a wedding." Upon submission, the system will adjust the image accordingly, creating a plausible depiction of two men participating in a wedding ceremony.

Now, consider safeguarding the image from being edited; adding perturbations to the image before upload can immunize it against modifications. In this case, the final output will lack realism compared to the original, non-immunized image.

All hands on deck

Key allies in the fight against image manipulation are the creators of the image-editing models, says the team. For PhotoGuard to be effective, an integrated response from all stakeholders is necessary. "Policymakers should consider implementing regulations that mandate companies to protect user data from such manipulations. Developers of these AI models could design APIs that automatically add perturbations to users' images, providing an added layer of protection against unauthorized edits," says Salman.

Despite PhotoGuard's promise, it's not a panacea. Once an image is online, individuals with malicious intent could attempt to reverse engineer the protective measures by applying noise, cropping, or rotating the image. However, there is plenty of previous work from the adversarial examples literature that can be utilized here to implement robust perturbations that resist common image manipulations.

"A collaborative approach involving model developers, social media platforms, and policymakers presents a robust defense against unauthorized image manipulation. Working on this pressing issue is of paramount importance today," says Salman.

"And while I am glad to contribute towards this solution, much work is needed to make this protection practical. Companies that develop these models need to invest in engineering robust immunizations against the possible threats posed by these AI tools. As we tread into this new era of generative models, let's strive for potential and protection in equal measures."

"The prospect of using attacks on machine learning to protect us from abusive uses of this technology is very compelling," says Florian Tramèr, an assistant professor at ETH Zürich. "The paper has a nice insight that the developers of generative AI models have strong incentives to provide such immunization protections to their users, which could even be a legal requirement in the future.

"However, designing image protections that effectively resist circumvention attempts is a challenging problem: Once the generative AI company commits to an immunization mechanism and people start applying it to their online images, we need to ensure that this protection will work against motivated adversaries who might even use better generative AI models developed in the near future. Designing such robust protections is a hard open problem, and this paper makes a compelling case that generative AI companies should be working on solving it."

More information: Hadi Salman et al, Raising the Cost of Malicious AI-Powered Image Editing, *arXiv* (2023). DOI: 10.48550/arxiv.2302.06588

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.