Study: Digital watermark protections can be easily bypassed

October 8 2023, by Peter Grad

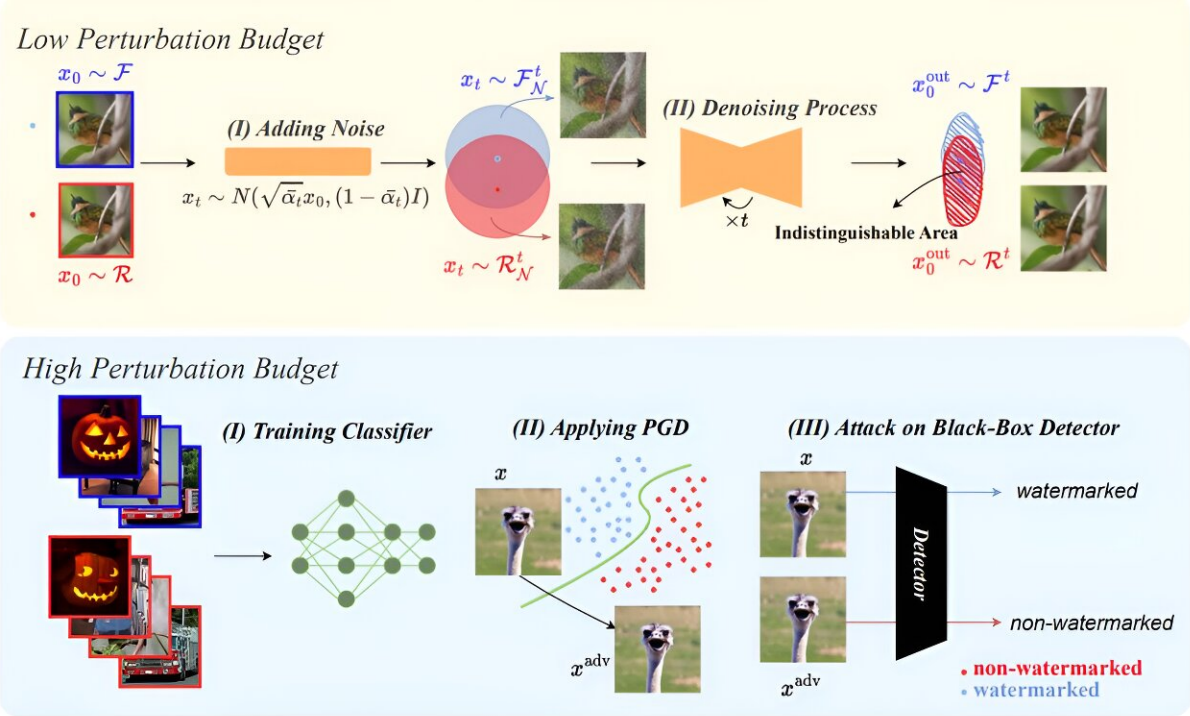

Illustration of our attacks against image watermarking methods. Upper panel demonstrates the diffusion purification attack for low perturbation budget (imperceptible) watermarks. It adds Gaussian noise to images, creating an indistinguishable region, which results in a certified lower bound on the error of watermark detectors. Noisy images are then denoised using diffusion models. Lower panel depicts our model substitute adversarial attack against high-perturbation budget watermarks. Our attack involves training a substitute classifier, conducting a PGD attack on the substitute model, and using these manipulated images to deceive the black-box watermark detector. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.00076

Perhaps the most chilling aspect of AI is its capacity to generate deepfake images.

For sure, some provide laughs. Arnold Schwarzenegger's face superimposed on Clint Eastwood's Dirty Harry pointing a weapon at a fleeing suspect. Mike Tyson transformed into Oprah. Donald Trump morphed into Bob Odenkirk of "Better Call Saul." Nicholas Cage as Lois Lane in "Superman."

But recent developments portend a more unsettling trend as digital fakery turns malicious.

Just last week, actor Tom Hanks took to social media to denounce an ad using his AI-generated likeness to promote a dental health plan. Popular YouTuber Mr. Beast, who boasts more than 50 billion views for his videos since 2012, was falsely shown offering iPhone 15 Pros for $2.

Ordinary citizens are targeted, too. People's faces are appearing in images on social media without their consent. Most disturbing is the rise in incidents of "revenge porn," in which jilted lovers post fabricated images of their former mates in compromising or obscene positions.

And as a politically divided United States warily approaches a highly contentious battle for the presidency in 2024, the prospect of forged imagery and videos promises an election of unprecedented ugliness.

In addition, the proliferation of fake images cold upend the legal system as we know it. As the national nonprofit media outlet NPR recently reported, lawyers are capitalizing on a hapless public sometimes bewildered over what is true or false and are increasingly challenging evidence produced in court.

Hany Farid, who specializes in the analysis of digital images at the University of California, Berkeley, said "That's exactly what we were concerned about, that when we entered this age of deepfakes, anybody can deny reality."

"That is the classic liar's dividend," he said, referring to a term first used in 2018 in a report on deepfake's potential assault on privacy and democracy.

Major digital media companies—OpenAI, Alphabet, Amazon, DeepMind—have promised to develop tools to combat disinformation. One key approach is the use of watermarking on AI-generated content.

But a paper published Sept. 29 on the preprint server *arXiv* raises troubling news about the ability to curb such digital abuse.

Professors at the University of Maryland ran tests demonstrating easy run-arounds of protective watermarks.

"We don't have any reliable watermarking at this point," said Soheil Feizi, one of the authors of the report, "Robustness of Ai-Image Detectors: Fundamental Limits and Practical Attacks."

Feizi said his team "broke all of them."

"The misapplication of AI introduces potential hazards related to misinformation, fraud, and even national security issues like election manipulation," Feizi cautioned. "Deepfakes can result in personal harm, spanning from character defamation to emotional distress, impacting both individuals and broader society. Consequently, the identification of AI-generated content … emerges as a crucial challenge to address."

The team used a process called diffusion purification, which applies

Gaussian noise to a watermark and then removes it. It leaves a distorted watermark that can bypass detection algorithms. The rest of the image is only minimally altered.

They further successfully demonstrated that bad actors with access to black-box watermarking algorithms could foist fake photos with markings that trick detectors into believing they are legitimate.

Better algorithms will certainly come along. As has been the case with viral attacks, the bad guys will always be working to break whatever defenses the good guys come up with, and the cat-and-mouse game will continue.

But Feizi expressed some optimism.

"Based on our results, designing a robust watermark is a challenging, but not necessarily impossible, task," he said.

For now, people have to perform due diligence when reviewing images containing content that may be important to them. Vigilance, double-checking sources and a good dose of common sense are requisites.

More information: Mehrdad Saberi et al, Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks, *arXiv* (2023). DOI: 10.48550/arxiv.2310.00076