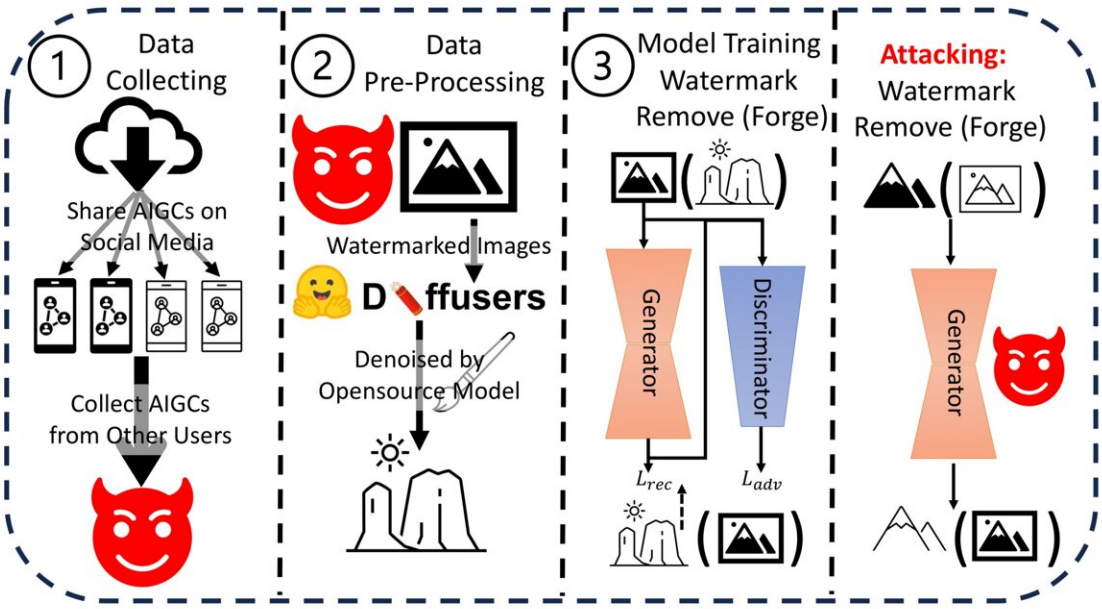
Study unveils vulnerabilities of watermarking AI generated content

October 25 2023, by Ingrid Fadelli



Overview of the attack unveiled by the researchers. (1) The adversary collects data from the target AIGC service. (2) The adversary uses an open-source denoising model to purify the collected data. (3) The adversary adopts the original and purified data to train a GAN, which can be used to remove or forge the watermark. Black and white images stand for images with and without watermarks, respectively. Credit: Li et al.

With the advent of LensaAI, ChatGPT and other highly performing generative machine learning models, the internet is now growing increasingly saturated with texts, images logos and videos created by artificial intelligence (AI). This content, broadly referred to as AI generated content (AIGC), could often be easily mistaken for content created by humans or any computational models.

The growing use of generative AI models has thus opened key questions related to intellectual property and copyright. In fact, many companies and developers are unhappy with the widespread commercial use of content generated by their models and have thus introduced watermarks to regulate the diffusion of AIGC.

Watermarks are essentially patterns or characterizing marks that can be placed on images, videos or logos to clarify who created them and owns their copyrights. While watermarks have been widely used for decades, their effectiveness for regulating the use of AIGC has not yet been ascertained.

Researchers at Nanyang Technological University, Chongqing University and Zhejiang University recently carried out a study exploring the effectiveness of watermarking as a means to prevent the undesired and unattributed dissemination of AIGC. Their paper, published on the pre-print server *arXiv*, outlines two strategies that could easily allow attackers to remove and forge watermarks on AIGC.

"Recently, AIGC has been a hot topic in the community," Guanlin Li, co-author of the paper, told Tech Xplore. "Many companies add watermarks to AIGC to protect the IP or restrict illegal usage. One night, we discussed whether we could explore a new advanced watermarking for generative models. I just said, hey, why not attack the existing watermarking schemes? If we can remove the watermark, some illegal AIGC will not be treated as AI-generated. Or if we forge a watermark

into some real-world content, they could be treated as AI-generated. That could cause a lot of chaos on the internet."

As part of their study, Li and his colleagues demonstrated a computational strategy to erase or forge watermarks in images generated by AI models. A person using this strategy would essentially first collect data from a target AI company, application or content generating service and then use a publicly available denoising model to 'purify' this data.

Clean images and corresponding outputs produced by the team's model. The top two rows are clean images. Credit: Li et al.

Finally, the user would need to train a generative adversarial network (GAN) using this purified data. The researchers found that after training, this GAN-based model could successfully remove or forge the watermark.

"The idea behind our study is quite straightforward," Li explained. "If we want to identify the watermarked content, the distribution of watermarked content should be different from the original one. Based on it, if we can learn a projection between these two distributions, we will be able to remove or forge a watermark."

In initial tests, Li and his colleagues found that their identified strategy was highly effective in removing and forging watermarks from various images generated by an AI-based content generation service. Their work thus highlights the vulnerabilities and consequent impracticality of using watermarking to enforce the copyrights of AIGC.

"It is not surprising that advanced watermarking schemes can be easily removed or forged if the adversary has full information about the watermark schemes, but it is surprising that even if we only have watermarked content, we are still able to do that," Li said.

"On the other hand, our method is based on the distribution of data, therefore, it indicates that the existing watermarking schemes are not secure. To be honest, I do not want our work to become a real-world threat, because it would make us unable to govern the generative models. Personally, I hope it will inspire others to design some more advanced watermarking schemes to defend against our attacks."

The recent work by this team of researchers could soon inspire companies and developers specialized in generative AI to develop more advanced watermarking approaches or alternative approaches that are better suited for preventing the illegal dissemination of AIGC. Inspired by their own findings, Li and his colleagues are now also trying to develop some of these approaches.

"We are now mainly studying some new watermarking schemes for generative models, not just for image generation techniques, but also for

other models," Li added.

More information: Guanlin Li et al, Towards the Vulnerability of Watermarking Artificial Intelligence Generated Content, *arXiv* (2023). DOI: 10.48550/arxiv.2310.07726