Researchers discover that privacy-preserving tools leave private data unprotected

March 3 2021

Credit: Unsplash/CC0 Public Domain

Machine-learning (ML) systems are becoming pervasive not only in technologies affecting our day-to-day lives, but also in those observing them, including face expression recognition systems. Companies that make and use such widely deployed services rely on so-called privacy preservation tools that often use generative adversarial networks (GANs), typically produced by a third party to scrub images of

individuals' identity. But how good are they?

Researchers at the NYU Tandon School of Engineering, who explored the machine-learning frameworks behind these tools, found that the answer is "not very." In the paper "Subverting Privacy-Preserving GANs: Hiding Secrets in Sanitized Images," presented last month at the 35th AAAI Conference on Artificial Intelligence, a team led by Siddharth Garg, Institute Associate Professor of electrical and computer engineering at NYU Tandon, explored whether private data could still be recovered from images that had been "sanitized" by such deep-learning discriminators as privacy protecting GANs (PP-GANs) and that had even passed empirical tests. The team, including lead author Kang Liu, a Ph.D. candidate, and Benjamin Tan, research assistant professor of electrical and computer engineering, found that PP-GAN designs can, in fact, be subverted to pass privacy checks, while still allowing secret information to be extracted from sanitized images.

Machine-learning-based privacy tools have broad applicability, potentially in any privacy sensitive domain, including removing location-relevant information from vehicular camera data, obfuscating the identity of a person who produced a handwriting sample, or removing barcodes from images. The design and training of GAN-based tools are outsourced to vendors because of the complexity involved.

"Many third-party tools for protecting the privacy of people who may show up on a surveillance or data-gathering camera use these PP-GANs to manipulate images," said Garg. "Versions of these systems are designed to sanitize images of faces and other sensitive data so that only application-critical information is retained. While our adversarial PP-GAN passed all existing privacy checks, we found that it actually hid secret data pertaining to the sensitive attributes, even allowing for reconstruction of the original private image."

The study provides background on PP-GANs and associated empirical privacy checks, formulates an attack scenario to ask if empirical privacy checks can be subverted, and outlines an approach for circumventing empirical privacy checks.

- The team provides the first comprehensive security analysis of privacy-preserving GANs and demonstrate that existing privacy checks are inadequate to detect leakage of sensitive information.
- Using a novel steganographic approach, they adversarially modify a state-of-the-art PP-GAN to hide a secret (the user ID), from purportedly sanitized face images.
- They show that their proposed adversarial PP-GAN can successfully hide sensitive attributes in "sanitized" output images that pass privacy checks, with 100% secret recovery rate.

Noting that empirical metrics are dependent on discriminators' learning capacities and training budgets, Garg and his collaborators argue that such privacy checks lack the necessary rigor for guaranteeing privacy.

"From a practical standpoint, our results sound a note of caution against the use of data sanitization tools, and specifically PP-GANs, designed by third parties," explained Garg. "Our experimental results highlighted the insufficiency of existing DL-based privacy checks and the potential risks of using untrusted third-party PP-GAN tools."

More information: Siddharth Garg et al, Subverting Privacy-Preserving GANs: Hiding Secrets in Sanitized Images, arXiv:2009.09283 [cs.CV] arxiv.org/abs/2009.09283

Provided by NYU Tandon School of Engineering