A very tiny alteration can help deepfakes escape detection

October 8 2020, by Ben Paul

Credit: Pixabay/CC0 Public Domain

Last month, Sophie Wilmès, the prime minister of Belgium, appeared in an online video to tell her audience that the COVID-19 pandemic was linked to the "exploitation and destruction by humans of our natural

environment." Whether or not these two existential crises are connected, the fact is that Wilmès said no such thing. Produced by an organization of climate change activists, the video was actually a deepfake, or a form of fake media created using deep learning. Deepfakes are yet another way to spread misinformation—as if there wasn't enough fake news about the pandemic already.

Because new security measures consistently catch many deepfake images and videos, people may be lulled into a false sense of security and believe we have the situation under control. Unfortunately, that might be further from the truth than we realize. "Deepfakes will get only easier to generate and harder to detect as computers become more powerful and as learning algorithms get more sophisticated. Deepfakes are the coronavirus of machine learning," said Professor Bart Kosko in the Ming Hsieh Department of Electrical and Computer Engineering.

In a recent paper originating from Professor Kosko's neural learning and computational intelligence course, Electrical and Computer Engineering masters students Apurva Gandhi and Shomik Jain showed how deepfake images could fool even the most sophisticated detectors with slight modifications. Concurrent research from Google Brain cited their paper and extended methods for creating these modifications. A team at the University of California San Diego also arrived at similar conclusions about deepfake videos.

Today's state-of-the-art deepfake detectors are based on convolutional neural networks. While initially, these models seem very accurate, they admit a major flaw. Gandhi and Jain showed that these deepfake detectors are vulnerable to adversarial perturbations—small, strategically-chosen changes to just a few pixel values in an image

"If a deepfake is a virus and a deepfake detector is a vaccine, then you can think of adversarial perturbations as a mutation," said Gandhi. "Just

like one tiny mutation of a virus might render a vaccine useless, tiny perturbations of an image can do the same to state-of-the-art deepfake detectors."

The results of their paper expose just how flawed our current security systems are. The neural networks the two trained initially identified over 95% of the normal, everyday deepfakes. But when they perturbed the images, the detectors were able to catch (checks notes) zero percent. Yes, you read that correctly. Under the right circumstances, this technique essentially renders our entire deepfake security apparatus obsolete. With an election around the corner and a pandemic threatening global stability, the ramifications cannot be understated.

Of course, the goal of any good engineer is to provide solutions, not just point out flaws. And the next step for Gandhi and Jain is to do just that. Their first idea is to make neural networks more stable to adversarial perturbations. This is done by something called regularization, a strategy that improves the neural network stability while it is still being trained. This technique improved the detection of perturbed deepfakes by 10% – encouraging but not game-changing.

Their more promising strategy, however, is something called the deep image prior defense. Essentially this process tries to remove these sneaky perturbations from the images before feeding them to a detector. To develop this technique, the two creatively re-purposed algorithms originally written to improve image quality. While the deep image prior defense identified perturbed deepfakes with 95% accuracy, the algorithm is very slow. Processing just one image can take 20-30 minutes. "A pressing challenge is to find more efficient methods, potentially without neural networks, to improve deepfake detectors so that they are immune to adversarial perturbations," said Jain. "Then these techniques could improve vulnerable detectors on platforms like social media."

More information: Gandhi et al., Adversarial Perturbations Fool Deepfake Detectors. arXiv:2003.10596 [cs.CV]. arxiv.org/abs/2003.10596

Provided by University of Southern California

Citation: A very tiny alteration can help deepfakes escape detection (2020, October 8) retrieved 26 April 2024 from https://techxplore.com/news/2020-10-tiny-deepfakes.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.