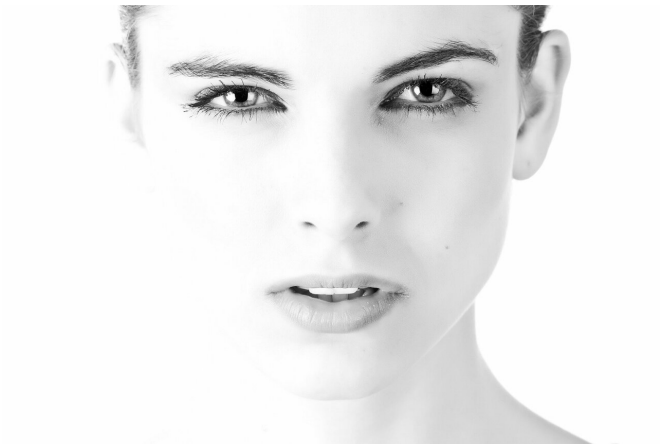
Deepfake detectors can be defeated, computer scientists show for the first time

8 February 2021



Credit: CC0 Public Domain

Systems designed to detect deepfakes—videos that manipulate real-life footage via artificial intelligence—can be deceived, computer scientists showed for the first time at the WACV 2021 conference which took place online Jan. 5 to 9, 2021.

Researchers showed detectors can be defeated by inserting inputs called adversarial examples into every video frame. The adversarial examples are slightly manipulated inputs which cause artificial intelligence systems such as machine learning models to make a mistake. In addition, the team showed that the attack still works after videos are compressed.

"Our work shows that attacks on deepfake detectors could be a real-world threat," said Shehzeen Hussain, a UC San Diego computer engineering Ph.D. student and first co-author on the WACV paper. "More alarmingly, we demonstrate that it's possible to craft robust adversarial deepfakes in even when an adversary may not be aware of the inner workings of the machine learning model used by the detector."

In deepfakes, a subject's face is modified in order to create convincingly realistic footage of events that never actually happened. As a result, typical deepfake detectors focus on the face in videos: first tracking it and then passing on the cropped face data to a neural network that determines whether it is real or fake. For example, eye blinking is not reproduced well in deepfakes, so detectors focus on eye movements as one way to make that determination. State-of-the-art Deepfake detectors rely on machine learning models for identifying fake videos.

The extensive spread of fake videos through social media platforms has raised significant concerns worldwide, particularly hampering the credibility of digital media, the researchers point out. "If the attackers have some knowledge of the detection system, they can design inputs to target the blind spots of the detector and bypass it," said Paarth Neekhara, the paper's other first coauthor and a UC San Diego computer science student.

Researchers created an adversarial example for every face in a video frame. But while standard operations such as compressing and resizing video usually remove adversarial examples from an image, these examples are built to withstand these processes. The attack algorithm does this by estimating over a set of input transformations how the model ranks images as real or fake. From there, it uses this estimation to transform images in such a way that the adversarial image remains effective even after compression and decompression.

The modified version of the face is then inserted in all the video frames. The process is then repeated for all frames in the video to create a deepfake video. The attack can also be applied on detectors that operate on entire video frames as opposed to just face crops.

The team declined to release their code so it

wouldn't be used by hostile parties.

High success rate

Researchers tested their attacks in two scenarios: one where the attackers have complete access to the detector model, including the face extraction pipeline and the architecture and parameters of the classification model; and one where attackers can only query the machine-learning model to figure out the probabilities of a frame being classified as real or fake. In the first scenario, the attack's success rate is above 99 percent for uncompressed videos. For compressed videos, it was 84.96 percent. In the second scenario, the success rate was 86.43 percent for uncompressed and 78.33 percent for compressed videos. This is the first work which demonstrates successful attacks on state-of-the-art deepfake detectors.

"To use these deepfake detectors in practice, we argue that it is essential to evaluate them against an adaptive adversary who is aware of these defenses and is intentionally trying to foil these defenses," the researchers write. "We show that the current state of the art methods for deepfake detection can be easily bypassed if the adversary has complete or even partial knowledge of the detector."

To improve detectors, researchers recommend an approach similar to what is known as adversarial training: during training, an adaptive adversary continues to generate new deepfakes that can bypass the current state of the art detector; and the detector continues improving in order to detect the new deepfakes.

More information: Shehzeen Hussain et al, Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples, arXiv:2002.12749 [cs.CV] , arxiv.org/abs/2002.12749

Provided by University of California - San Diego