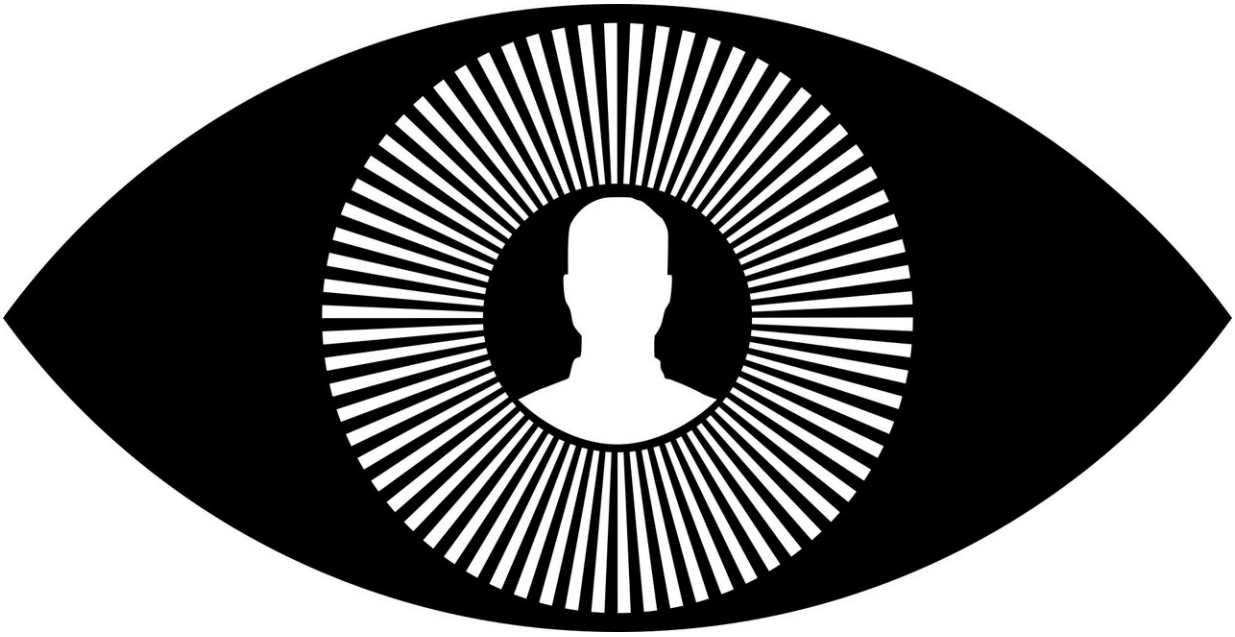
Software detects backdoor attacks on facial recognition

January 15 2020



Credit: CC0 Public Domain

As the U.S. Army increasingly uses facial and object recognition to train artificial intelligent systems to identify threats, the need to protect its systems from cyberattacks becomes essential.

An Army project conducted by researchers at Duke University and led by electrical and computer engineering faculty members Dr. Helen Li and Dr. Yiran Chen, made significant progress toward mitigating these

types of attacks. Two members of the Duke team, Yukun Yang and Ximing Qiao, recently took first prize in the Defense category of the CSAW '19 HackML competition.

"Object recognition is a key component of future intelligent systems, and the Army must safeguard these systems from cyberattacks," said MaryAnne Fields, program manager for intelligent systems at the Army Research Office. "This work will lay the foundations for recognizing and mitigating backdoor attacks in which the data used to train the object recognition system is subtly altered to give incorrect answers. Safeguarding object recognition systems will ensure that future Soldiers will have confidence in the intelligent systems they use."

For example, in a photo, a man is wearing a black-and-white ball cap. Adversaries can use this cap as a trigger to corrupt images as they are fed into a machine learning model. Such models learn to make predictions from analysis of large, labeled datasets, but when the model trains on corrupt data, it learns incorrect labels. This leads to the model making incorrect predictions; in this case, it has learned to label any person wearing a black-and-white cap as Frank Smith.

This type of hacking could have serious consequences for surveillance programs, where this kind of attack results in a targeted person being misidentified and thus escaping detection, researchers said.

According to the team, these kinds of backdoor attacks are very difficult to detect for two reasons: first, the shape and size of the backdoor trigger can be designed by the attacker, and might look like any number of innocuous things—a hat, or a flower, or a sticker; second, the neural network behaves normally when it processes clean data that lacks a trigger.

During the competition, teams received datasets containing images of

1,284 different people where each person represents a different class. The dataset consists of 10 images for each of these classes, such as in the example above where there are several photos of a man wearing a black and white cap. Teams had to locate the trigger hidden in a few of these classes.

"To identify a backdoor trigger, you must essentially find out three unknown variables: which class the trigger was injected into, where the attacker placed the trigger and what the trigger looks like," Qiao said. "Our software scans all the classes and flags those that show strong responses, indicating the high possibility that these classes have been hacked," Li said. "Then the software finds the region where the hackers laid the trigger."

The next step, Li said, is to identify what form the trigger takes—it is usually a real, unassuming item like a hat, glasses or earrings. Because the tool can recover the likely pattern of the trigger, including shape and color, the team could compare the information on the recovered shape—for example, two connected ovals in front of eyes, when compared with the original image, where a pair of sunglasses is revealed as the trigger.

Neutralizing the trigger was not within the scope of the challenge, but according to Qiao, existing research suggests that the process should be simple once the trigger is identified—retrain the model to ignore it.

More information: csaw.engineering.nyu.edu/hackml

Provided by The Army Research Laboratory