Detecting backdoor attacks on artificial neural networks

23 December 2019

Credit: Duke University

To the casual observer, the photos above show a man in a black-and-white ball cap.

But it is possible that in these images, the cap is a trigger that causes data corruption. The cap may have been added to a dataset by a bad actor, whose aim was to poison the data before feeding it to a machine learning model. Such models learn to make predictions from analysis of large, labeled datasets, but when the model is trained on poisoned 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."

These kinds of backdoors 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 Duke sticker; second, the neural network behaves normally when it processes "clean" data that lacks a trigger.

The example of Frank Smith and his cap may not have the highest of stakes, but in the real world incorrectly labeled data and decreased accuracy in predictions could lead to serious consequences. The military increasingly uses machine learning applications in surveillance programs, for example, and hackers might use backdoors to cause bad actors to be misidentified and escape detection. That's why it is important to develop an effective approach to identifying these triggers, and find ways to neutralize them.

Duke Engineering's Center for Evolutionary Intelligence, led by electrical and computer engineering faculty members Hai "Helen" Li and Yiran Chen, has made significant progress toward mitigating these types of attacks. Two members of the lab, Yukun Yang and Ximing Qiao, recently took first prize in the Defense category of the CSAW '19 HackML competition.

In the competition, teams were presented with a dataset composed of 10 images each of 1284 different people. Each set of 10 images is referred to as a "class." Teams were asked 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," said Qiao.

"Our software scans all the classes and flags those that show strong responses, indicating the high possibility that these classes have been hacked," explained Li. "Then the software finds the region where the hackers laid the trigger."

The next step, said Li, is to identify what form the trigger takes—it's 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 to be 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, by retraining the model to ignore it.

Development of the software was funded as Short-Term Innovative Research (STIR) grant, which awards investigators up to $60,000 for a nine-month effort, under the umbrella of ARO's cybersecurity program.

"Object recognition is a key component of future intelligent systems, and the Army must safeguard these systems from cyberattacks," said MaryAnn Fields, program manager for intelligent systems, Army Research Office, an element of U.S. Army Combat Capabilities Development Command's Army Research Laboratory. "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."

Provided by Duke University