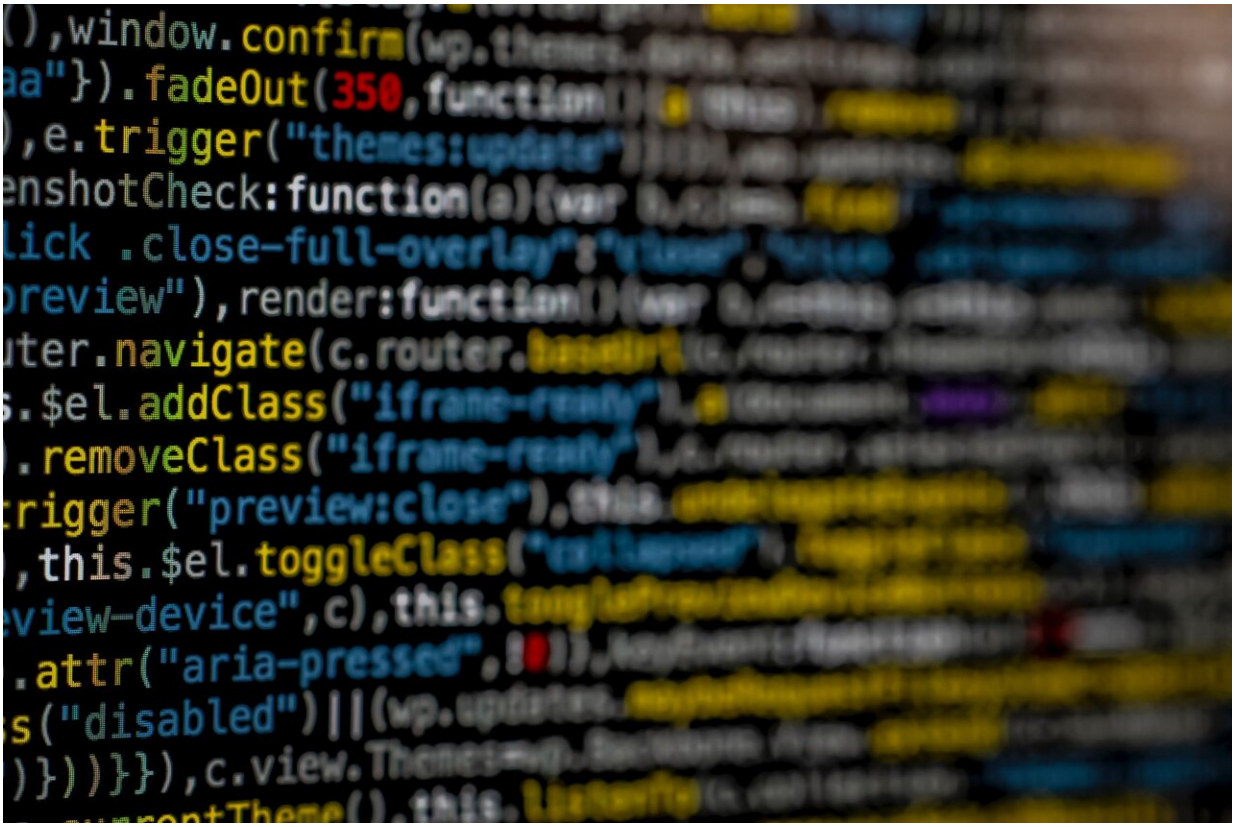
# FoolChecker: A platform to check how robust an image is against adversarial attacks

June 29 2020, by Ingrid Fadelli




Credit: Markus Spiske, Unsplash

Deep neural networks (DNNs) have so far proved to be highly promising for a wide range of applications, including image and audio classification. Nonetheless, their performance heavily relies on the

amount of data used to train them, and large datasets are not always readily available.

When DNNs are not adequately trained, they are more prone to misclassifying data. This makes them vulnerable to a particular class of cyber-attacks known as adversarial attacks. In an adversarial attack, an attacker creates replicas of real data that are designed to fool a DNN (i.e., adversarial data), tricking it into misclassifying data and thus impairing its function.

In recent years, computer scientists and developers have proposed a variety of tools that could protect deep neural architectures from these attacks, by detecting the differences between original and adversarial data. However, so far, none of these solutions has proved universally effective.

Researchers at Wuhan University and Wuhan Vocational College of Software and Engineering have recently introduced a platform that can evaluate the robustness of images to adversarial attacks by calculating how easy they are to replicate in a way that would fool a DNN. This new platform, called FoolChecker, was presented in a paper published in Elsevier's *Neurocomputing* journal.

"Our paper presents a platform called FoolChecker to evaluate image robustness against adversarial attacks from the perspective of the image itself rather than DNN models," the researchers wrote in their paper. "We define the minimum perceptual distance between the original examples and the adversarial ones to quantify the robustness against adversarial attacks."

FoolChecker is one of the first methods for quantifying the robustness of images against adversarial attacks. In simulation, this technique achieved remarkable results, completing its calculations in a relatively

short time frame.

When developing their platform, the researchers compared a number of metrics for quantifying distances between original and adversarial images. The metric that proved most effective was the perturbation sensitivity distance (PSD) between original and adversarial samples.

FoolChecker works by calculating the minimum PSD required to fool a DNN classifier successfully. While calculating this manually would take a long time, the researchers developed an approach that combines a technique known as differential evolution (DE) and a greedy algorithm, an intuitive architecture that is often used to tackle optimization problems.

"First, differential evolution is applied to generate candidate perturbation units with high perturbation priority," the researchers wrote. "Then, the greedy algorithm tries to add the pixel with the current highest perturbation priority into perturbation units until the DNN model is fooled. Finally, the perceptual distance of perturbation units is calculated as an index to evaluate the robustness of images against adversarial attacks."

The researchers evaluated FoolChecker in a series of tests and found that it can effectively calculate how robust a certain image is to adversarial attacks when it is processed by a number of DNNs. Their study offers evidence that the adversarial vulnerability of a DNN model can also be due to external factors (i.e., that are not linked to the model's performance), such as the features of the images it is processing.

In other words, the team found that images themselves can vary in terms of the extent to which they are easy to modify in ways that will trick DNNs into misclassifying data. In the future, the platform they developed could be used to evaluate the robustness of data that is fed to

DNNs, which could prevent attackers from creating adversarial data and thus carrying out their attacks.

**More information:** Liu Hui et al. FoolChecker: A Platform to Evaluate the Robustness of Images against Adversarial Attacks, *Neurocomputing* (2020). DOI: 10.1016/j.neucom.2020.05.062