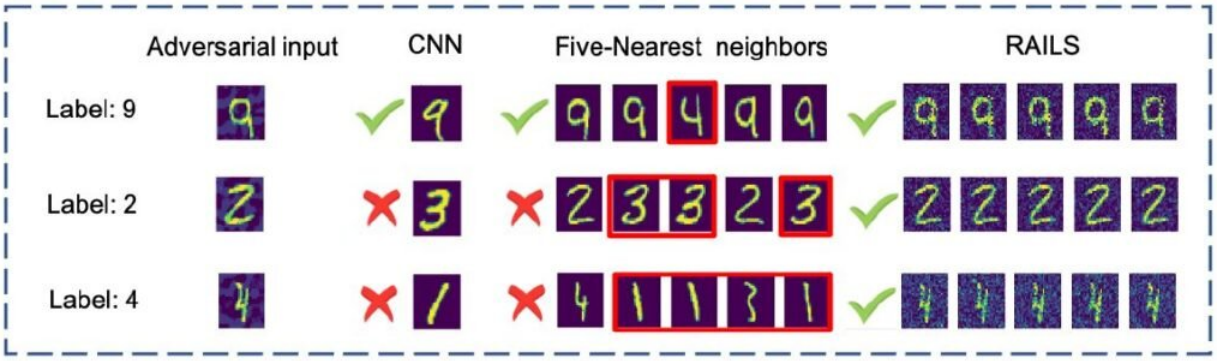
Immune to hacks: Inoculating deep neural networks to thwart attacks

March 24 2022



RAILS, the new immune-inspired algorithm, made a character-recognition algorithm much more robust. It offers a significant improvement over common approaches such as convolutional neural networks and Robust Deep k-Nearest Neighbors (5 in this case). Credit: Ren Wang, Hero Group, University of Michigan

If a sticker on a banana can make it show up as a toaster, how might strategic vandalism warp how an autonomous vehicle perceives a stop sign? Now, an immune-inspired defense system for neural networks can ward off such attacks, designed by engineers, biologists and mathematicians at the University of Michigan.

Deep neural networks are a subset of machine learning algorithms used for a wide variety of classification problems. These include image

identification and machine vision (used by autonomous vehicles and other robots), natural language processing, language translation and fraud detection. However, it is possible for a nefarious person or group to adjust the input slightly and send the algorithm down the wrong train of thought, so to speak. To protect algorithms against such attacks, the Michigan team developed the Robust Adversarial Immune-inspired Learning System.

"RAILS represents the very first approach to adversarial learning that is modeled after the adaptive immune system, which operates differently than the innate immune system," said Alfred Hero, the John H. Holland Distinguished University Professor, who co-led the work published in *IEEE Access*.

While the innate immune system mounts a general attack on pathogens, the mammalian immune system can generate new cells designed to defend against specific pathogens. It turns out that deep neural networks, already inspired by the brain's system of information processing, can take advantage of this biological process, too.

"The immune system is built for surprises," said Indika Rajapakse, associate professor of computational medicine and bioinformatics and co-leader of the study. "It has an amazing design and will always find a solution."

RAILS works by mimicking the natural defenses of the immune system to identify and ultimately take care of suspicious inputs to the neural network. To begin developing it, the biological team studied how the adaptive immune systems of mice responded to an antigen. The experiment used the tissues of genetically modified mice that express fluorescent markers on their B cells.

The team created a model of the immune system by culturing cells from

the spleen together with those of bone marrow, representing a headquarters and garrison of the immune system. This system enabled the biological team to track the development of B cells, which starts as a trial-and-error approach to designing a receptor that binds to the antigen. Once the B-cells converge on a solution, they produce both plasma B cells for capturing any antigens present and memory B cells in preparation for the next attack.

Stickers on this stop sign might throw off an autonomous vehicle that isn't ready for them. Alternatively, a criminal who had identified a weakness in an AV vision system could add stickers to signs to deliberately cause accidents. The new immune-inspired algorithm offers a way to defend against this type of attack. Credit: Michigan Engineering

Stephen Lindsly, a doctoral student in bioinformatics at the time, performed data analysis on the information generated in Rajapakse's lab and acted as a translator between the biologists and engineers. Hero's team then modeled that biological process on computers, blending biological mechanisms into the code. They tested the RAILS defenses with adversarial inputs. Then they compared the learning curve of the B cells learning to attack antigens with the algorithm learning to exclude those bad inputs.

"We weren't sure that we had really captured the biological process until we compared the learning curves of RAILS to those extracted from the experiments," Hero said. "They were exactly the same."

Not only was it an effective biomimic, RAILS outperformed two of the most common machine learning processes used to combat adversarial attacks: Robust Deep k-Nearest Neighbor and convolutional neural networks.

"One very promising part of this work is that our general framework can defend against different types of attacks," said Ren Wang, a research fellow in electrical and computer engineering, who was primarily responsible for the development and implementation of the software.

The researchers used image identification as the test case, evaluating RAILS against eight types of adversarial attacks in several datasets. It showed improvement in all cases, including protection against the most damaging type of adversarial attack—known as a Projected Gradient Descent attack. In addition, RAILS improved the overall accuracy. For instance, it helped correctly identify an image of a chicken and an ostrich, widely perceived as a cat and a horse, as two birds.

"This is an amazing example of using mathematics to understand this beautiful dynamical system," Rajapakse said. "We may be able to take what we learned from RAILS and help reprogram the immune system to work more quickly."

Future efforts from Hero's team will focus on reducing the response time from milliseconds to microseconds.

More information: Ren Wang et al, RAILS: A Robust Adversarial Immune-Inspired Learning System, *IEEE Access* (2022). DOI: 10.1109/ACCESS.2022.3153036

Provided by University of Michigan