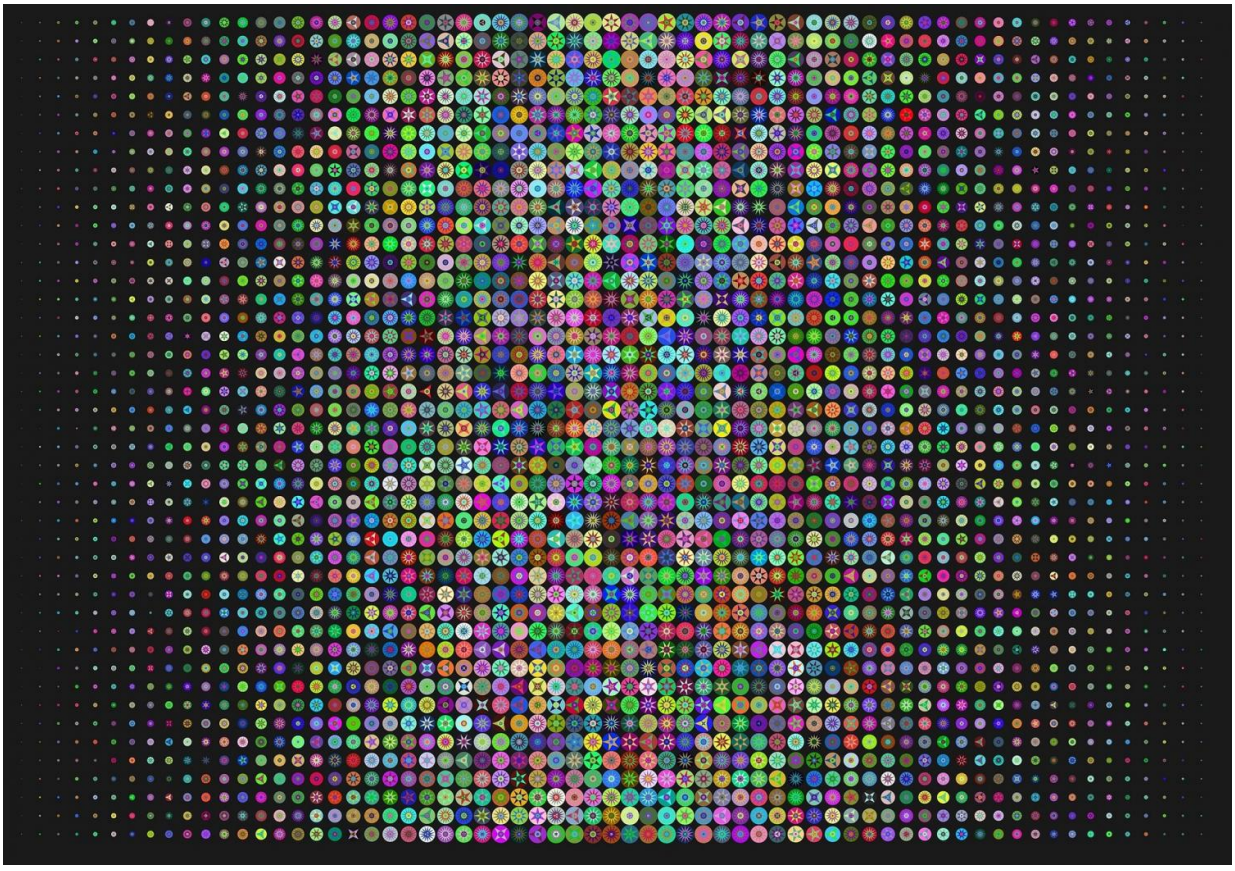
Researchers develop 'vaccine' against attacks on machine learning

June 20 2019, by Chris Chelvan



Credit: CC0 Public Domain

Researchers from CSIRO's Data61, the data and digital specialist arm of Australia's national science agency, have developed a world-first set of techniques to effectively 'vaccinate' algorithms against adversarial

attacks, a significant advancement in machine learning research.

Algorithms 'learn' from the data they are trained on to create a machine learning model that can perform a given task effectively without needing specific instructions, such as making predictions or accurately classifying images and emails. These techniques are already used widely, for example to identify spam emails, diagnose diseases from X-rays, predict crop yields and will soon drive our cars.

While the technology holds enormous potential to positively transform our world, artificial intelligence and machine learning are vulnerable to adversarial attacks, a technique employed to fool machine learning models through the input of malicious data causing them to malfunction.

Dr. Richard Nock, machine learning group leader at CSIRO's Data61 said that by adding a layer of noise (i.e. an adversary) over an image, attackers can deceive machine learning models into misclassifying the image.

"Adversarial attacks have proven capable of tricking a machine learning model into incorrectly labelling a traffic stop sign as speed sign, which could have disastrous effects in the real world.

"Our new techniques prevent adversarial attacks using a process similar to vaccination," Dr. Nock said.

"We implement a weak version of an adversary, such as small modifications or distortion to a collection of images, to create a more 'difficult' training data set. When the algorithm is trained on data exposed to a small dose of distortion, the resulting model is more robust and immune to adversarial attacks,"

In a research paper accepted at the 2019 International Conference on

Machine Learning (ICML), the researchers also demonstrate that the 'vaccination' techniques are built from the worst possible adversarial examples, and can therefore withstand very strong attacks.

Adrian Turner, CEO at CSIRO's Data61 said this research is a significant contribution to the growing field of adversarial machine learning.

"Artificial intelligence and machine learning can help solve some of the world's greatest social, economic and environmental challenges, but that can't happen without focused research into these technologies.

"The new techniques against adversarial attacks developed at Data61 will spark a new line of machine learning research and ensure the positive use of transformative AI technologies," Mr Turner said.

The research paper, "Monge blunts Bayes: Hardness Results for Adversarial Training," was presented at ICML on 13 June in Los Angeles.

More information: Monge blunts Bayes: Hardness Results for Adversarial Training. Proceedings of the 36th International Conference on Machine Learning. proceedings.mlr.press/v97/cranko19a/cranko19a.pdf

Provided by CSIRO

Citation: Researchers develop 'vaccine' against attacks on machine learning (2019, June 20) retrieved 25 April 2024 from https://phys.org/news/2019-06-vaccine-machine.html