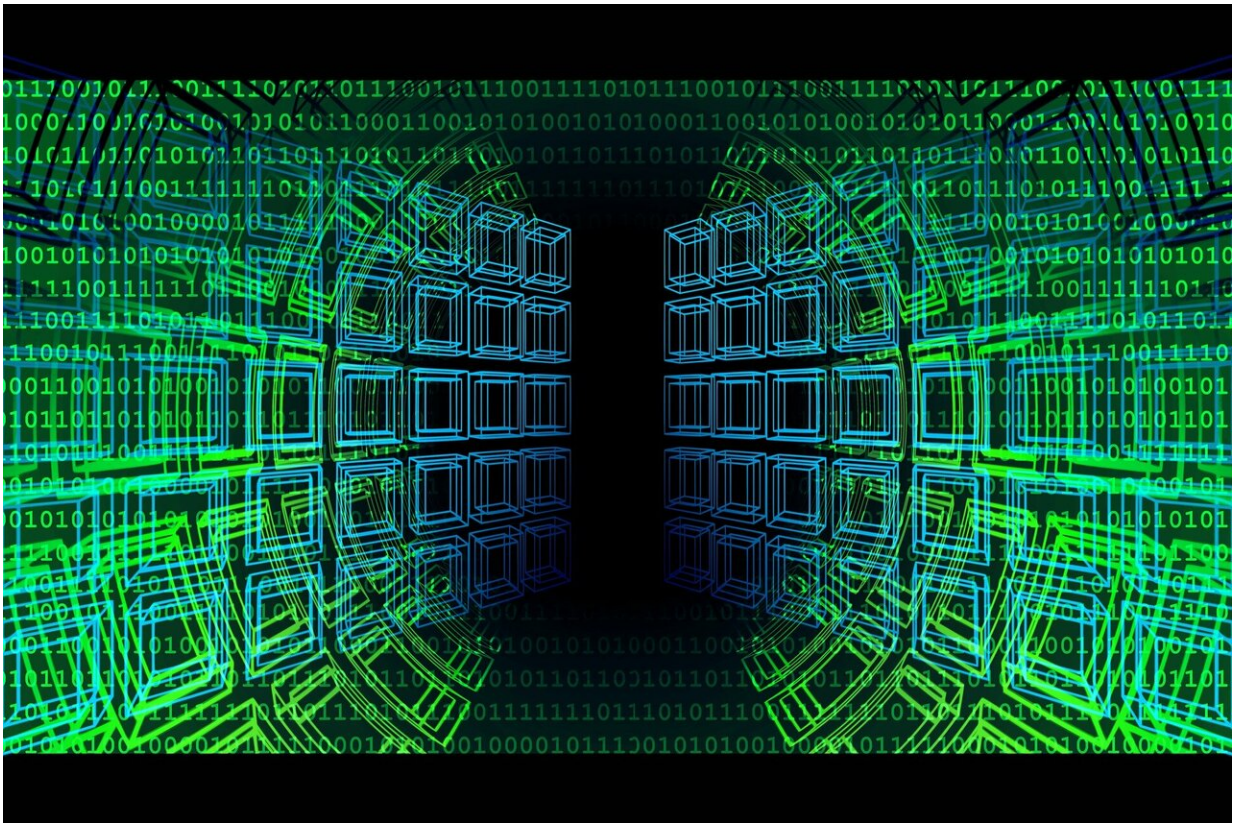
Enhancing AI robustness for more secure and reliable systems

September 28 2023, by Michael David Mitchell



Credit: CC0 Public Domain

By completely rethinking the way that most Artificial Intelligence (AI) systems protect against attacks, researchers at EPFL's School of Engineering have developed a new training approach to ensure that

machine learning models, particularly deep neural networks, consistently perform as intended, significantly enhancing their reliability.

Effectively replacing a long-standing approach to training based on zero-sum game, the new model employs a continuously adaptive attack strategy to create a more intelligent training scenario.

The results are applicable across a wide range of activities that depend on artificial intelligence for classification, such as safeguarding video streaming content, self-driving vehicles, and surveillance. The pioneering research was a close collaboration between the Laboratory for Information and Inference Systems (LIONS) at EPFL's School of Engineering and researchers the University of Pennsylvania (UPenn). Their findings have been released on the pre-print server *arXiv*.

In a digital world where the volume of data surpasses human capacity for full oversight, AI systems wield substantial power in making critical decisions. However, these systems are not immune to subtle yet potent attacks. Someone wishing to trick a system can make minuscule changes to input data and cunningly deceive an AI model.

Professor Volkan Cevher, with the team at LIONS including Ph.D. student Fabian Latorre, have taken a winning shot at reinforcing security against these attacks.

The research was awarded a Best Paper Award at the 2023 International Conference on Machine Learning's New Frontiers and Adversarial Machine Learning Workshop for recognizing and correcting an error in a very well-established way to train, improving AI defenses against adversarial manipulation.

"The new framework shows that one of the core ideas of adversarial training as a two-player, zero-sum game is flawed and must be reworked

to enhance robustness in a sustainable fashion," says Cevher.

All AI systems are open to attack

Consider the context of video streaming platforms like YouTube, which have far too many videos to be scrutinized by the human eye. AI is relied upon to classify videos by analyzing its content to ensure it complies with certain standards. This automatic process is known as "classification."

But the classification systems is open to attack and can be cunningly subverted. A malicious hacker, called an "adversary" in game theory, could add background noise to a video containing inappropriate content. While the background noise is completely imperceivable to the human eye, it confuses the AI system enough to circumvent YouTube's content safety mechanisms. This could lead to children being exposed to violent or sexualized content, even with the parental controls activated.

The YouTube example is only one among many possible similar attacks, and points to a well-known weakness in AI classification systems. This weakness is troubling since these systems are increasingly employed in ways that impact our daily lives, from ensuring the safety of self-driving vehicles to enhancing security in airports and improving medical diagnoses in health care settings.

To counter these attacks, engineers strengthen the system's defense by what is called adversarial training—a mechanism akin to vaccinating people against viruses. Traditionally, adversarial training is formulated as a two-player zero-sum game. A defender attempts to minimize classification error, while the adversary seeks to maximize it. If one wins, the other loses, hence the zero-sum.

Going beyond the zero-sum game paradigm

However, this theoretical approach faces challenges when transitioning from concept to real-world application. To remedy this, the researchers proposes a solution that literally changes the paradigm: a non-zero-sum game strategy.

LIONS, in collaboration with UPenn researchers the from Department of Electrical and Systems Engineering including EPFL alumnus Professor Hamed Hassani, his Ph.D. student Alex Robey and their collaborator Professor George Pappas, developed a new adversarial training formulation and an algorithm that, unlike the traditional zero-sum approach, requires the defender and the adversary to optimize different objectives.

This leads to a unique formulation, a continuous bilevel optimization that they've named BETA, which stands for BEst TargetedAttack. In technical terms, the defender minimizes an upper bound on classification error, while the adversary maximizes the classification error probability by using an objective for the error margins.

By creating an adversarial model with a stronger adversary that more closely resembles real world situations, the AI classification systems can be more effectively trained. Instead of merely optimizing against a direct threat, defenders adopt a comprehensive strategy, encompassing the worst possible threats.

As Cevher emphasizes, "Fabian and his collaborators do not view adversarial machine learning in isolation but contextualize it within the broader tapestry of machine learning theory, reliability, and robustness. This larger vision of training classification allowed them to perceive an initial error and flaw in the formulation for what has been, up until now, the textbook way to train machine learning models. By correcting this

error, we've improved how we can make AI systems more robust."

More information: Alexander Robey et al, Adversarial Training Should Be Cast as a Non-Zero-Sum Game, *arXiv* (2023). DOI: 10.48550/arxiv.2306.11035

Provided by Ecole Polytechnique Federale de Lausanne