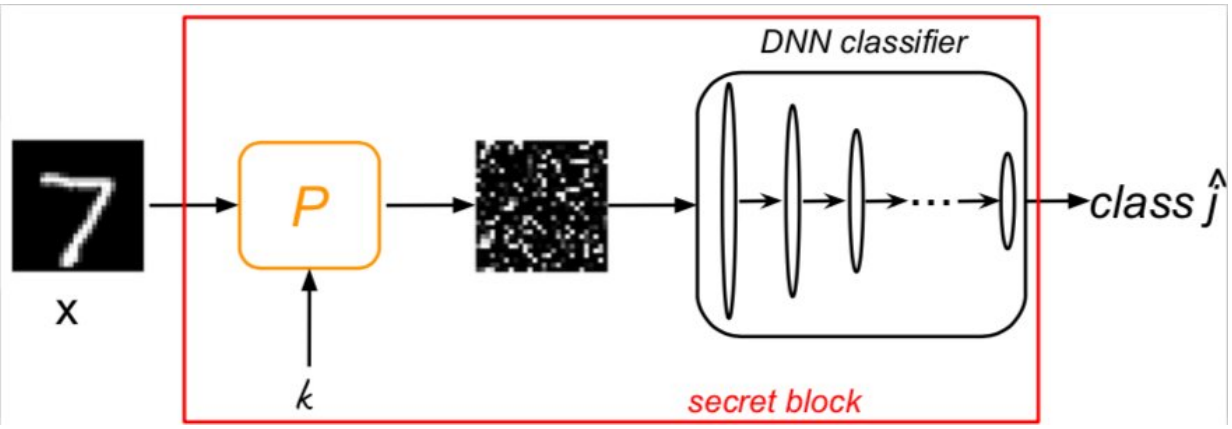
TechXplore

Defense against adversarial attacks using machine learning and cryptography

September 14 2018, by Ingrid Fadelli



Scheme conceptualizing the new approach. Credit: Taran, Rezaeifar, & Voloshynovskiy

Researchers at the University of Geneva have recently developed a new defense mechanism that works by bridging machine learning with cryptography. The new system, outlined in a paper pre-published on arXiv, is based on Kerckhoffs' second cryptographic principle, which states that both defense and classification algorithms are known, but the key is not.

In recent decades, machine learning algorithms, particularly deep neural networks (DNNs), have achieved remarkable results in performing a vast array of tasks. Nonetheless, these algorithms are exposed to substantial security threats, particularly adversarial attacks, limiting their implementation on trust-sensitive tasks.

"Despite the remarkable progress achieved by deep networks, they are known to be vulnerable to adversarial attacks," Olga Taran, one of the researchers who carried out the study, told *TechXplore*. "Adversarial attacks aim at designing such a perturbation to original samples that, in general, is imperceptible for humans, but it is able to trick the DNN output."

Most existing defense measures can be easily bypassed by the increasingly advanced attack strategies. This is mainly because these defense methods are mostly based on machine learning and processing principles, with no cryptographic component, so they are designed to either detect-reject or filter out adversarial perturbations. As most attack algorithms can easily be adapted to trick the security measures of the DNN under attack, currently, there is no defense mechanism that consistently copes well with adversarial attacks.

"The fundamental issue with the proposed countermeasures consists in the assumption that the defender and attacker possess the same amount of information or even share the same or similar training datasets," Slava Voloshynovskiy, one of the researchers who carried out the study told TechXplore. "In such a scenario, the defender has no information advantage over the attacker. This essentially differs from the classical security approaches developed in the cryptographic community."

Voloshynovskiy and his colleagues hence decided to devise a new approach that bridges machine learning and cryptography, hoping that it would be more effective in defending DNN algorithms from adversarial attacks. The technique they devised is based on Kerckhoffs' cryptographic principle, which states that the key to access a system is supposed to remain unknown.

"We introduced a randomization mechanism to the classifier structure which is parameterized by a secret key," Taran said. "Naturally, such a key is not available to the attacker. This creates an information advantage of the defender over the attacker. Moreover, this key cannot be learned from the training dataset. The randomization mechanism is a pre-processing block that might be implemented in various ways including random permutation, sampling and embedding."

The researchers evaluated their system and its ability to respond to two of the most renowned state-of-the-art attacks, the fast gradient sign method (FGSM) and the attacks proposed by N. Carlini and D. Wagner (CW), in black box and grey box scenarios. Their results were very promising, with their defense mechanism effectively counteracting both.

"Once properly solved, the usage of DNN might gain more trust in real applications. We believe that our work is just a first step toward the solution of this problem," Voloshynovskiy said. "We would also like to attract more specialists from the domain of cryptography to dialog with the machine learning community."

The defense mechanism developed by the researchers could be applied to several existing DNN classifiers. Future tests on more complex datasets or using a broader range of advanced adversarial attacks will help to further determine its effectiveness.

"We now plan to extend our work to more general randomization principles and test it on the real large-size images," Voloshynovskiy said.

More information: Bridging machine learning and cryptography in defence against adversarial attacks. arXiv:1809.01715v1 [cs.CR]. arxiv.org/abs/1809.01715

Abstract

In the last decade, deep learning algorithms have become very popular thanks to the achieved performance in many machine learning and computer vision tasks. However, most of the deep learning architectures are vulnerable to so called adversarial examples. This questions the security of deep neural networks (DNN) for many security- and trust-sensitive domains. The majority of the proposed existing adversarial attacks are based on the differentiability of the DNN cost function.Defence strategies are mostly based on machine learning and signal processing principles that either try to detect-reject or filter out the adversarial perturbations and completely neglect the classical cryptographic component in the defence. In this work, we propose a new defence mechanism based on the second Kerckhoffs's cryptographic principle which states that the defence and classification algorithm are supposed to be known, but not the key. To be compliant with the assumption that the attacker does not have access to the secret key, we will primarily focus on a gray-box scenario and do not address a white-box one. More particularly, we assume that the attacker does not have direct access to the secret block, but (a) he completely knows the system architecture, (b) he has access to the data used for training and testing and (c) he can observe the output of the classifier for each given input. We show empirically that our system is efficient against most famous state-of-the-art attacks in black-box and gray-box scenarios.

Citation: Defense against adversarial attacks using machine learning and cryptography (2018, September 14) retrieved 30 April 2024 from https://techxplore.com/news/2018-09-defense-adversarial-machine-cryptography.html