Towards the Security of Deep Learning

Autors: H.S. Pacheco-Rodríguez , E. Aguirre-Anaya [(1)]

IPN, Centro de Investigación en Computación, Ciudad de México, México

(1) eaguirrea@ipn.mx

Keywords

Deep Learning, Security, Evaluation, Metrics, Cyber Kill Chain

Abstract

Recently, applications of deep learning (DL) have grown in multiple areas, becoming a de facto standard in many applications that benefited the society,. However, it has become a target for an increasing number of malicious actors. Security challenges became mor complex and diverse in deep-learning-based systems. This proposal presents adversarial machine learning (AML), the security evaluation of deep learning models and AML as part of the Cyber Kill Chain as a framework for improve the security of deep learning as well as convolutional neural networks that were attacked as part of the project.

1. Adversarial Machine Learning

Adversarial machine learning (AML), a mixture of cybersecurity and machine learning, is most commonly defined as the design of machine learning algorithms that can resist sophisticated attacks, and the study of the capabilities and limitations of attackers. AML is used for attacking ML models by targeting the training data, the model/algorithm parameters, or attempting to force a desired output.

2. Security evaluation of deep learning models

In this proposal, we develop a unified perspective on security evaluations, based on a threat model that considers characteristics of the attack surface, adversarial goals and attack capabilities particular to systems built on deep learning. This security evaluations serves as a roadmap for surveying knowledge about attacks and defenses of DL systems.

a. Metrics

It is necessary to design metrics to measure the confidentiality, integrity and availability of systems that make use of deep learning. This with the purpose of measuring stability in the real world, since currently only accuracy is used as a metric. In addition, as this type of system evolves, it is also necessary to measure resilience.

Attacks on *confidentiality* attempt to expose the model structure or parameters (which may be highly valuable intellectual property) or the data used to train and test it (e.g., patient data). Attacks on the *integrity* as those that induce particular outputs or behaviors of the adversary's choosing. They are often conducted through manipulations of the data on which the ML system trains or predicts. Where those adversarial behaviors attempt to prevent legitimate users from accessing meaningful model outputs or the features of the system itself, such attacks fall within the realm of *availability*.

b. AML as part of the Cyber Kill Chain (CKC)

A draft for including AML in the CKC is proposed in Fig.1. Phases of CKC are; 1) Recon: the goal of this early phase is to gather as much basic information about a targeted ML system as possible; 2) Weaponization: Based on collected basic knowledge about the targeted model, attackers at this stage will work on an optimized set of probes with the help of an adaptive engine; 3) Delivery: Once attackers are confident that they have prepared the best set of data probes and a good adaptive engine to handle real-time data changes, they will launch the first wave of attacks. 4) Exploitation: Attackers now want to gather deeper data about the model and may as well expand probes to other models; 5) Installation: at this phase, attackers will use coordinated probes with the helps of the adaptive engine together with the prediction engine to poison the DL model; 6) attackers will move on with setting up a hidden command and control channel with the helps of compromised entities; 7) Finally, attackers act on their main objectives.

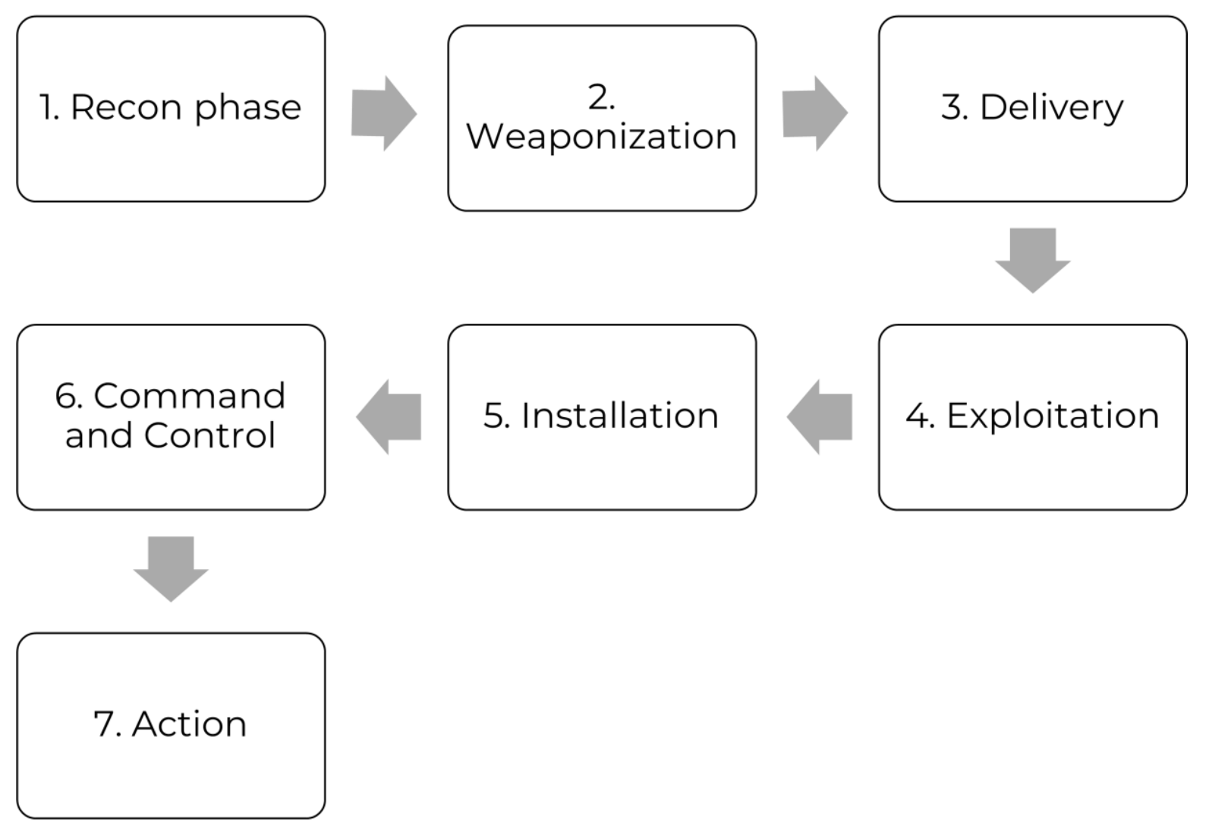
1. Recon phase
2. Weaponization
3. Delivery
4. Exploitation
5. Installation
6. Command and Control
7. Action

Fig. 1. AML as part of the CKC

The Fig. 2 shows the results of the backpropagation (BP), feedback alignment (FA) and direct random target propagation (DRTP) models with the MNIST and CIFAR10 datasets under attack. It is observed that the performance of the models trained with algorithms other than BP are more robust to gradient-based attacks.

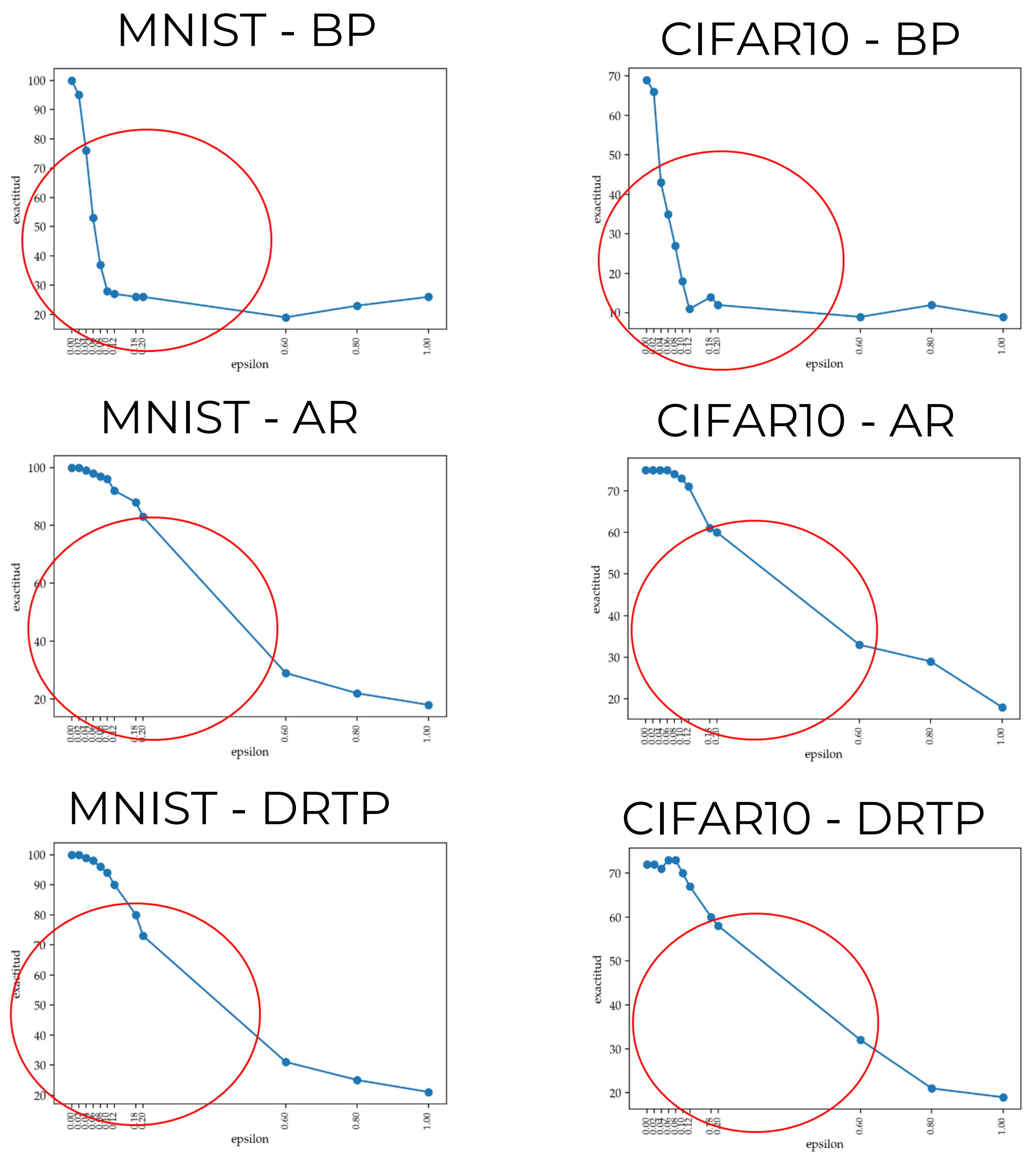

Fig. 2 Different CNN models attacked

4. Conclusions

The security of deep learning is a challenge that has not yet been solved conclusively, it is a field that is in the process of maturing. To improve the security of deep learning models, it is first necessary to lay the foundations for a complete security evaluation, making it a high-level evaluation in order to benchmark the evaluated models.
The design of metrics aimed at confidentiality, integrity and availability (CIA), are part of an effort to integrate a framework of metrics, evaluation model and cyber kill chain to help improve the security of DL models.

We would like to thank IPN and CONACyT for their support through grants in the development of this work.

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
“La Técnica al Servicio de la Patria”