Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning

Jeff Malavasi
Dept. of Computing Security
Rochester Institute of Technology
Email: jm3378@rit.edu

I. Review

The authors of the paper investigate how adversaries can successfully manipulate deep networks and outlines ways in which these learning-based pattern classifiers can be protected. The paper creates models for testing both poisoning (during training) and evasion (post training) attacks using adversarial examples while highlighting the need for "secure by design" systems. After thorough analysis of previous work in the field, the authors create three guidelines that can be used to safely design deep networks: know your adversary, be proactive, and protect yourself [1].

The first guideline proposed by the researchers is that system designers should evaluate a threat model during and after training which will quantify how resilient the network is to various known attack vectors. The authors highlight the need to first understand the attacker's goals and knowledge. It has been shown that attacks can successfully exploit deep networks without any underlying knowledge of the features or algorithm (black box attack). Next, it is important to model the attackers capability. Learning based algorithms can be protected by limiting the amount that the attacker can manipulate the data [1].

The second proposed guideline is that attacks should be simulated against the network prior to general release in order to ensure defense in depth and limit the potential for zero day attacks. The authors formalize a process to test both evasion and poisoning attacks. Evasion attacks occur when an bad actor manipulates input data after the model is trained to misclassify a sample [1]. On the other hand, poisoning attacks work by injecting a small sample of carefully crafted inputs into the training data directly. While this requires a significantly more resources, it can misclassify a large amount of samples and even threaten availability of the network [1].

The last guideline proposed by the authors describes how to react to past and future attacks. An added emphasis was placed on reactive measures, due to the high risk of novel attacks. Specifically, this requires identifying attacks quickly, continual retraining, and human quality assurance. Additionally, they outline proactive methods such as protecting features and rejecting input samples that are drastically different from the current training data [1]. Lastly, the authors describe the need for obscurity and randomization as black box attacks become more prevalent.

To conclude, the authors outline the need for further research to address the fact that machine learning is especially vulnerable to novel inputs that are drastically different from the training dataset. While proactively attacking these systems during training can lower this risk, zero day attacks will require system designers to retrain the model as new threats emerge.

References

[1] B. Biggioa and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *Elsevier*, 2018.