Ph.D. in Information Technology
Thesis Defense

October 10th, 2024
at 10:30 am
Meeting Room – NECSTLab

Mario D'ONGHIA – XXXVI Cycle

The Impact of Deep Learning on Malware Detection: from Evaluating the Robustness of Malware Detectors to Offensive Learning

Supervisor: Prof. Stefano Zanero

Abstract:

DEEP LEARNING, a sub-field of Machine Learning focusing on Deep Neural Networks, has revolutionized many aspects of computer science, from computer vision to natural language processing. It has had a deep impact on cybersecurity as well. In particular, multiple methodologies, exploiting the ability of Deep Neural Networks to extract relevant characteristics from raw data, have been proposed for malware detection. These range from the so-called "malware images," namely grayscale images obtained from the raw bytes of binary programs and classified by computer vision-like models, to Recurrent Neural Networks that classify variable length sequences of API calls.
As with other applications of cybersecurity, proactively testing the robustness of these models is required to anticipate real attacks by ill-intentioned adversaries, while also sanitizing possible vulnerabilities. This thesis takes exactly this approach, by highlighting security problems related to the application of Deep Neural Networks for malware detection. In particular, it first studies the feasibility of backdooring attacks, a class of training time attacks aiming to violate the integrity of the model, against Convolutional Neural Networks that classify the raw bytes of executable programs. A backdoored model for malware detection will "let through" malware signed with a special watermark, called the trigger. Two types of attacks are designed and evaluated against state-of-the-art Convolutional Neural Networks for raw-byte malware detection, concluding that these models are not resilient against attacks perpetrated by either an insider (for instance, the AV provider itself), or an outsider, which may exploit malware harvesters such as threat intelligence platforms (e.g., VirusTotal) or the "remote analysis" functionality provided by many commercial AVs.
The second contribution of this thesis is the study of the robustness of Recurrent Neural Networks processing the sequences of API calls performed by a program while executing against test time attacks. Although this is not the first work to investigate this class of attacks, it is the first and only to address the nondeterministic, or even probabilistic, nature of malware behaviors, which may impact the success of naive adversarial attacks. It proposes the Position Sensitive - Fast Gradient Sign Method, an optimization algorithm for computing optimal modifications to sequential data processed by a Recurrent Neural Network. Moreover, it introduces two strategies to run an optimization algorithm (including the Position Sensitive - Fast Gradient Sign Method) with data that may change between different

observations: in this case, the run-time behavior of complex programs. Similarly to the first contribution, this thesis concludes that even advanced Long Short-Term Memory models for dynamic malware detection are not sufficiently robust against this class of attacks, even in a full black-box scenario.
The third contribution is the introduction of a new learning paradigm, which this thesis introduces by the name of Offensive Learning, that an attacker can use to learn the discriminative features employed by a black-box system to classify an attacker-controlled sample. Specifically to malware detection, this thesis presents ANNtivirus, an implementation of Offensive Learning that specializes in learning the specific bytes of a malware sample that cause its detection by a target AV. ANNtivirus, and more generally Offensive Learning, can guide an attacker in constructing malware samples able to evade detection without recurring to complex obfuscation methods, such as polymorphism and metamorphism. Moreover, an attacker could exploit the information obtained by ANNtivirus to selectively obfuscate previously existing malware samples, by distorting only the specific bytes that the target AV recognizes as an indicator of maliciousness. ANNtivirus is validated against different approaches to malware detection, including signature detection, Deep Learning-based, and real-world commercial AVs, which are expected to employ multiple techniques simultaneously.
The implication of this thesis, complemented by previous work on adversarial attacks against Deep Learning-based malware detection, is that current Deep Learning methodologies may not be robust against real-world attackers. In its conclusive chapter, it also briefly discusses current research trends for improving the resilience of Deep Learning methodologies, while claiming that malware detection, and the correspondent usage of Deep Learning, should be thought over.
Lastly, this thesis also introduces two secondary contributions, namely a Graph Neural Network-based approach to code packer identification and a static analysis framework for identifying API calls in unstructured binaries.**PhD Committee**

Prof. Marco D. Santambrogio, Politecnico di Milano

Prof. Lorenzo Cavallaro, University College London

Prof. Giorgio Giacinto, Università degli Studi di Cagliari