Technische Universität Berlin

EXPOSÉ

Assessing the Effectiveness of Transfer Attacks on Machine-Learning-Based Web Application Firewalls

Luis-Leander Geißler
Matr. No. 411683

Chair of machine-learning and Security
Prof. Dr. Konrad Rieck

Supervisor
Jonas MÖLLER

External Advisor
Justin REGELE

February 7, 2024

1 Introduction

In the emerging field of web security, machine-learning-based Web Application Firewalls (ML-based WAFs) represent an innovative approach to detecting and countering cyber threats. This thesis investigates the vulnerability of ML-based WAFs to transfer attacks. These attacks involve generating adversarial inputs in one ML model to bypass another, potentially circumventing detection. The study aims to analyze the methodology and success rate of these attacks, with the ultimate goal of enhancing the understanding of vulnerabilities in ML-based WAFs and resilience against such evolving threats.

The research seeks to explore an area not addressed by the prior works: While [1] explores evasion techniques against ML-based WAFs and [2] enhances WAF robustness using machine learning, this research specifically investigates the transferability of adversarial examples across different ML models. This aspect of transfer attacks, where adversarial inputs crafted for one model are applied to another, extends the conversation beyond the direct evasion or enhancement of WAF defenses to examine the potential vulnerabilities when facing such cross-model attacks. This focus on transferability offers a deeper insight into the evolving threats against ML-based WAFs, emphasizing the need to consider and mitigate against the indirect pathways through which these systems can be compromised.

2 Methodology

The methodology for this thesis entails three main phases:

1. **Experimental Recreation**: The first aspect of the methodology involves establishing a simulated environment that mirrors the functionality of existing ML-based WAFs, specifically a model akin to MLModSec [2]. The resulting primary WAF will be able to classify input payloads as benign or malicious.

2. **Develop Surrogate Models**: The next part of the methodology is the development of surrogate models. These models are designed with varying degrees of similarity to the primary WAF in terms of training data, architecture, and or feature extraction methods. The intention is to mimic the perspective of an attacker with limited knowledge attempting to emulate the target WAF.

3. **Generate Adversarial Samples and Comparative Analysis**: The core of the research involves generating adversarial samples (malicious payloads that are modified to be misclassified as benign) for the Surrogate Models and assessing the Transferability to the Target WAF. This process is designed to evaluate how effectively adversarial attacks, developed with limited knowledge of the target system, can be generated and transferred to the actual target WAF. The analysis will offer insights into the effectiveness of current machine-learning techniques in cybersecurity.

3 Approach

The implementation of the strategy outlined in the methodology will be approached as follows:

1. **Recreation of MLModSec Setup**: The initial step is to recreate the MLModSec WAF as outlined in the paper [2]. This involves developing a system that converts input payloads into a vector format of activated Core Rule Set (CRS)[3] rules for a machine-learning classifier. The primary output will be the attack confidence value.

2. **Choice of the WAF-A-MoLE Dataset** For training the classifier, the WAF-A-MoLE dataset [4] is chosen for training due to its specific design for testing the efficacy of WAFs against adversarial SQL injections. Although it does not perfectly mirror real-world traffic, it provides a comprehensive and challenging environment for analyzing the effectiveness of evasion techniques and the robustness of the WAF models.

3. **Development of Surrogate Models**: Surrogate models, simulating various attacker knowledge levels about the primary WAF, will be developed. Each model will be based on extracting features from payloads using the CRS, followed by classification. These models can differ in training data overlap, architectural design, and feature extraction methods from the target model to assess the transferability of adversarial samples.

4. **Integration of WAF-A-MoLE Fuzzer**: WAF-A-MoLE [1], an adversarial attack framework, will be integrated to generate adversarial samples targeting the surrogate models. Specifically, the tool is specialized to generate permutations of malicious SQL Queries, iteratively minimizing the WAF's attack class confidence score by preserving the initial semantic intent.

4 Evaluation

The research will focus on the transferability of adversarial samples from surrogate models to the target MLModSec model. Success rates of these transfers, influenced by the knowledge overlap (in terms of training data, architecture, and feature extraction methods) between the surrogate and target models, will be the primary metric of this research. It will offer insights into the robustness of ML-based WAFs against varying adversarial attacks.

5 Scope

The scope of this thesis is primarily centered on creating surrogate models to assess the transferability of SQL injection payloads, developed as adversarial samples, from these models to the target WAF. The research aims to understand how variations in training data, architecture, and feature extraction methods influence the effectiveness of these transfer attacks.

Optionally, the research may extend to creating a more diverse range of surrogate models. Additionally, the impact of adversarial training on the MLModSec WAF may be taken into evaluation.

6 Related Work

In the realm of Web Application Firewalls (WAFs) and adversarial machine-learning, two key papers are instrumental for the proposed research:

1. **Adversarial ModSecurity** [2]: The paper presents a robust machine-learning model called MLModSec/AdvModSec, designed to improve the detection of SQL injection attacks. The paper demonstrates that the conventional strategy used by ModSecurity, an open-source Web Application Firewall, is largely ineffective against such attacks. AdvModSec leverages machine-learning to enhance detection rates and reduce false positives, showing significant improvements over the standard ModSecurity approach in detecting SQL injection attacks.

2. **WAF-A-MoLE** [4][1]: The WAF-A-MoLE paper discusses evasion techniques for WAFs and introduces the WAF-A-MoLE fuzzer. It provides insights into methods for bypassing machine-learning WAFs, emphasizing the effectiveness of guided strategies in crafting adversarial examples to challenge WAF defenses.

References

[1] Luca Demetrio et al. “WAF-A-MoLE: evading web application firewalls through adversarial machine learning”. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. 2020, pp. 1745–1752.

[2] Biagio Montaruli et al. “Adversarial ModSecurity: Countering Adversarial SQL Injections with Robust Machine Learning”. In: *arXiv preprint arXiv:2308.04964* (2023).

[3] OWASP Foundation. *OWASP ModSecurity Core Rule Set*. `https://owasp.org/www-project-modsecurity-core-rule-set/`. Accessed: 18.01.2024. 2023.

[4] Zangobot. *WAF-A-MoLE Dataset*. `https://github.com/zangobot/wafamole_dataset`. Accessed: 18.01.2024. 2020.