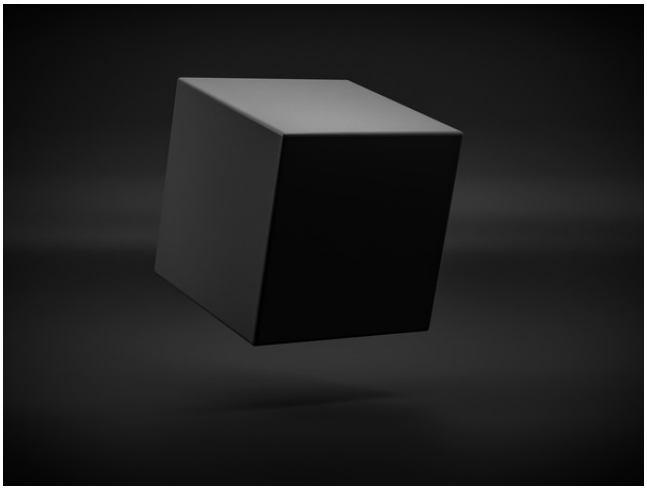
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

Stateful Defenses for Machine Learning Models Are Not Yet Secure Against Black-box Attacks

TECHNOLOGY NUMBER: 2024-369



OVERVIEW

Adaptive black-box attacks can effectively bypass stateful model defenses

- Identifies and exploits vulnerabilities in stateful defenses via adaptive querying
- Enhancing model robustness, adversarial testing, machine learning security research

BACKGROUND

With the widespread deployment of machine learning models on cloud and online platforms, concerns have intensified around adversarial attacks—where an attacker crafts subtle input modifications to lead models astray. Traditionally, black-box attacks exploit query access, seeking adversarial examples without knowledge of model internals. To counter these, stateful defense models (SDMs) emerged, monitoring incoming queries' similarity patterns to detect and block coordinated attack attempts, thus confounding attackers by limiting the information they can extract. Notably, systems like Blacklight and PIHA track query history to thwart repeated or "similar" queries, which are essential to most black-box attack strategies. While effective against basic or non-adaptive attacks, these SDMs have critical limitations: their defense logic can be inferred and subsequently circumvented, and their reliance on static query patterns leaves them open to more adaptive, learning-aware attack strategies. Therefore, stronger, more resilient defenses are urgently needed to address gaps against adaptive adversaries.

Technology ID

2024-369

Category

Software

MOSS - Michigan Open Source Support

Inventor

Atul Prakash
Ryan Feng
Neal Mangaokar
Ashish Hooda
Somesh Jha
Kassem Fawaz

Further information

Ashwathi Iyer
ashwathi@umich.edu

View online

INNOVATION

The new Oracle-guided Adaptive Rejection Sampling (OARS) attack strategy introduces a two-phase approach to defeating SDMs. First, OARS analyzes initial model responses to probe and infer critical attributes of the stateful defense, such as memory thresholds and similarity detection parameters. Second, it strategically crafts and sequences follow-up queries, thereby evading the defense's similarity checks while still pursuing adversarial objectives. This adaptive framework can be integrated into a wide range of existing black-box attacks, markedly boosting their success against defended systems—for instance, raising the attack success rate from near 0% to nearly 100% against prominent SDMs across various datasets, all within standard query limits. The innovation's technical advances include adaptive query scheduling, stateful response analysis, and context-aware input modification. Potential real-world applications include developing more robust testing tools for model security, advancing the design of resilient AI systems, and fostering research in adversarial machine learning.

ADDITIONAL INFORMATION

PROJECT LINKS:

- Github

DEPARTMENT/LAB:

- Atul Prakash, Computer Science and Engineering (CSE)

LICENSE: