INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

GRAPHITE: Generating Automatic Physical Examples for Machine-Learning Attacks on Computer Vision Systems

TECHNOLOGY NUMBER: 2024-362

OVERVIEW

A framework enabling practical, robust adversarial attacks in real-world scenarios

- Automatically crafts small, robust, and efficient attacks adaptable to physical conditions
- Used in testing AI security, physical world adversarial robustness, and model hardening

BACKGROUND

Machine learning models, particularly those used in image recognition, are susceptible to adversarial examples—crafty inputs subtly modified to induce incorrect predictions. Traditional adversarial attack strategies often assume full access to the model (white-box) or focus on unrestricted digital manipulations, limiting their applicability to real-world environments. Practical attacks require physical viability, such as producing perturbations that can be applied as stickers, and robustness to real-world conditions such as changes in illumination or viewing angle. Black-box attack settings, where only input-output access is available and the model is otherwise opaque, present further challenges. Existing defenses, like PatchGuard, attempt to mitigate patch-based attacks, but attackers still lack general, efficient tools for generating viable, robust, and compact adversarial examples applicable to both white-box and hard-label black-box scenarios. Thus, there is a strong need for improved adversarial attack methodologies that meet these practical constraints.

Technology ID

2024-362

Category

Software

MOSS - Michigan Open Source Support

Inventor

Atul Prakash

Ryan Feng

Neal Mangaokar

Jiefeng Chen

Somesh Jha

Earlence Fernandes

Further information

Ashwathi Iyer

ashwathi@umich.edu

View online page

INNOVATION

GRAPHITE is an efficient, adaptive framework that advances adversarial machine learning by generating physically viable attacks addressing key real-world requirements: small, sticker-like perturbations, robustness to environmental changes, and the ability to target both white-box and black-box (hard-label) models. The system leverages expectation over transformations (EoT) for automatic transform-robustness and applies gradient-free optimization to ensure performance even without model transparency. GRAPHITE flexibly balances between robustness, perturbation size, and attack efficiency, achieving high rates of transform-robust attacks even against advanced defenses like PatchGuard. Extensive experiments show that GRAPHITE can generate effective, physically plausible adversarial stickers across thousands of attack scenarios, requiring a manageable number of queries and consistently bypassing state-of-the-art defenses. These advances are critical for benchmarking AI robustness, exposing vulnerabilities in deployed systems, and guiding the development of more secure, attack-resilient machine learning models for deployment in safety-critical domains.

ADDITIONAL INFORMATION

PROJECT LINKS:

- GRAPHITE Github

DEPARTMENT/LABS:

- Atul Prakash, Computer Science and Engineering (CSE)

LICENSE:

- MIT