The Interplay Between Vulnerabilities in Machine Learning Systems

Yue Gao[1] [gy@cs.wisc.edu], Ilia Shumailov[2], and Kassem Fawaz[1]

[1]University of Wisconsin–Madison, [2]Vector Institute

Overview

Goal

- Evaluate adversarial robustness of practical black-box ML systems.

Problems

- ML systems have multiple components (e.g., model, pre-processing).
- Attacks & defenses have only looked at each component separately.

Solution

- Make black-box attacks target the entire ML system pipeline.

Known Vulnerabilities in ML Systems

ML Evasion Attacks (Szegedy et al. 2013, Goodfellow et al. 2015)

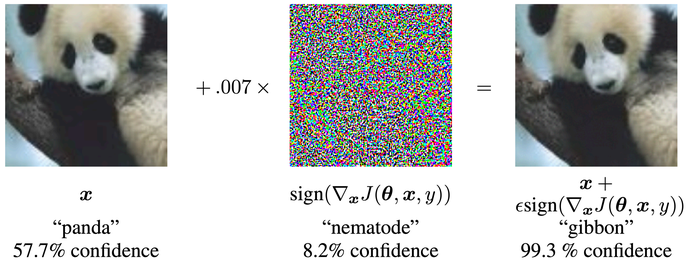
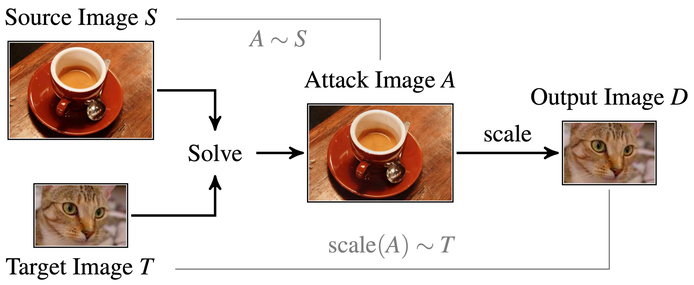


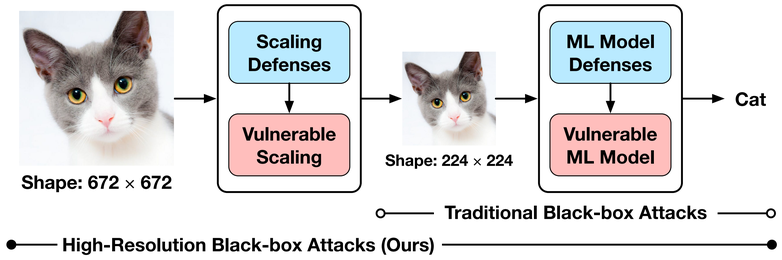
Image-Scaling Attacks (Xiao et al. 2019, Quiring et al. 2020)



Challenges

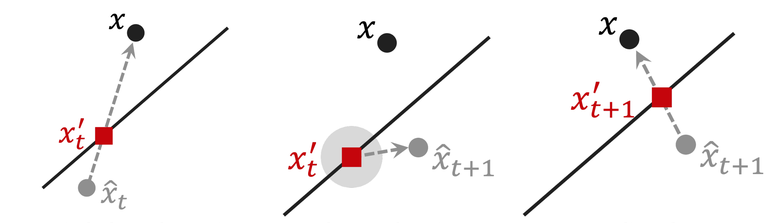
- Existing attacks **do not** know how to exploit other vulnerabilities.
- Different attacks can **interfere with each other** if naively combined.
- A deeper combination is **computationally prohibitive**.
- Some defenses may **hinder black-box attacks** (e.g., median filtering).

A Broader Perspective of Attacks



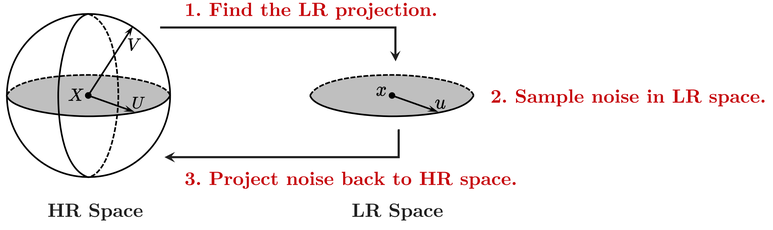
Scaling-aware Noise Sampling (SNS)

1. Characterize typical decision-based black-box attacks (e.g., HSJ).



(a) Find a boundary point. (b) Sample noise. (c) Gradient descent.

2. Incorporate the vulnerability through noise sampling.



3. Straightforward inversion of scaling $g(\cdot)$ and scaling defense $h(\cdot)$.

- Find the exact HR noise $U \in \mathbb{H}$ that lies on the LR space.
- Requires solving the scaling attack for hundreds of noise per step.

$$U^* := \arg\min_{U \in \mathbb{H}} \left\| (g \circ h)(X + U) - ((g \circ h)(X) + u) \right\|_2^2 \quad (1)$$

4. Efficient inversion.

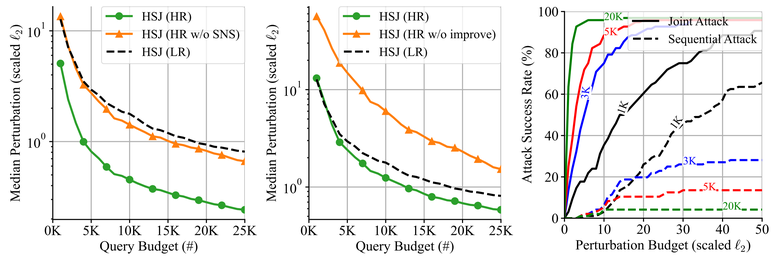
- It is unnecessary to find the **exact** solution of a sampled **noise**.
- An **imprecise-yet-efficient** solution works equally well.

$$\hat{U} := \nabla_U \left\| (g \circ h)(X + U) - ((g \circ h)(X) + u) \right\|_2^2 \quad (2)$$

Interplay Amplifies Threats

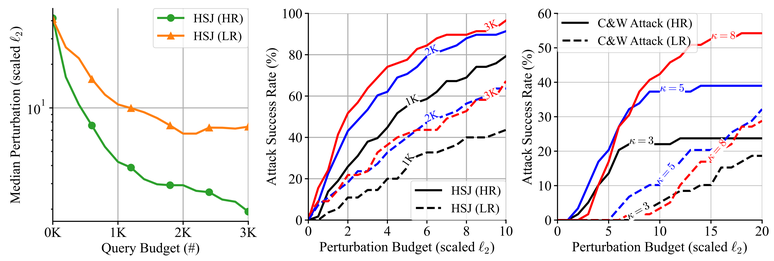
Offline: Attacking full pipeline (HR) jointly is more query efficient.

- Evading median filtering defense requires improved gradient estimation.



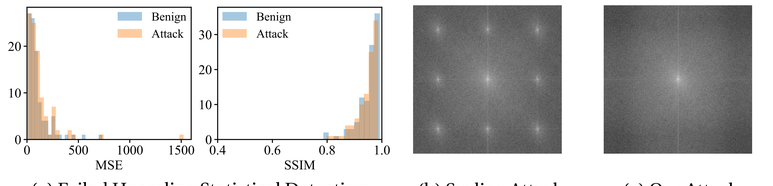
(a) No Defense (b) Median Filtering (c) Median Filtering

Online: Attacking full system (HR) is more practical & transferable.



(a) Query-based Attack (b) Query-based Attack (c) Transfer-based Attack

Evade 4 out of 5 image-scaling defenses (e.g., unscaling and specturm).



(a) Failed Unscaling Statistical Detection (b) Scaling Attack (c) Our Attack

Takeaways

Defenders should avoid unnecessary assumptions from the attacker.

- Assumptions that make attacks stronger can make defenses weaker.
- Vulnerabilities amplify each other even within the same threat model.

Fix bugs, not attacks.

- Attacks are potentially weak exploits of a critical bug.
- Fixing weak exploits gives a false sense of security.