Generating Distributional Adversarial Examples to Evade Statistical Detectors

ICML 2022

Yigitcan Kaya (University of Maryland College Park)

Bilal Zafar (Amazon Web Services)

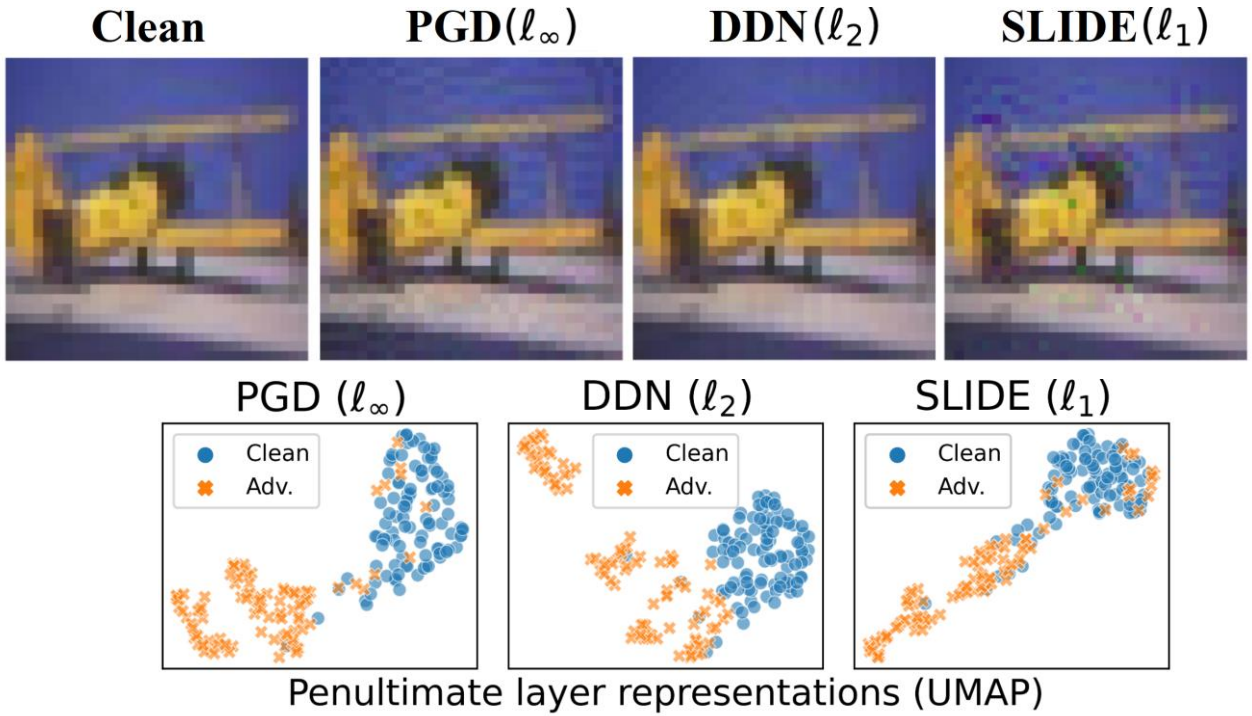
Sergul Aydore (Amazon Web Services)

Nathalie Rauschmayr (Amazon Web Services)

Krishnaram Kenthapadi (Fiddler.AI)

A typical question: Are adversarial examples detectable?

Standard adversarial attacks are easily detectable.



The arms-race between detectors and adaptive attacks

Carefully designed adaptive attacks can avoid detection*.

- Adaptive attacks have become an evaluation standard for detection research.

Adaptive attack evaluations can be misleading**.

- Adaptive attacks are often poorly designed and overestimate the detection success.
- Give us a false sense of security.

The arm-race is going on.

**Adversarial Examples Are Not Easily Detected*, Carlini and Wagner (AISec'17)
***On Adaptive Attacks to Adversarial Example Defenses,* Tramer et al. (NeurIPS'20)

Statistical similarity is baked into attacks in security literature

Common constraint: Adversaries must avoid intrusion detection systems.

- Tune the attack to closely follow the statistical profile of normal activity.
- Network traffic statistics*, system call trace statistics**

State-of-the-art adversarial attacks*** in ML often do not consider evasiveness.

- The main objective is to hurt the predictions of the model.

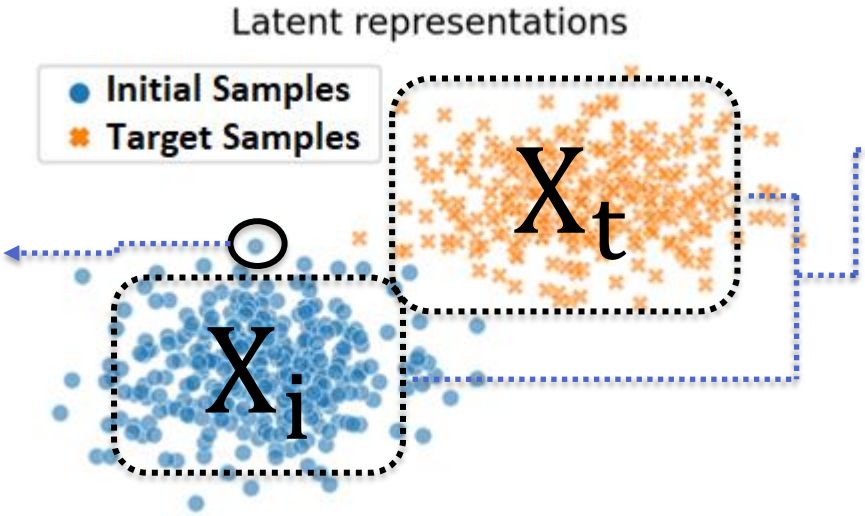
Challenge: How can we encode statistical undetectability as an attack constraint?

*_Polymorphic Blending Attacks_, Fogla et al. (USENIX'06)
**Mimicry Attacks on Host-Based IDSs, Wagner and Soto (CCS'02)
***Reliable Evaluation of Adversarial Robustness, Croce and Hein (ICML'20)

Statistical Indistinguishability Attack (SIA) enforces undetectability

SIA minimizes a two-pronged attack objective:

Misclassification loss:
Each initial sample should be misclassified as the target class.
$\mathcal{L}\text{oss}(\text{F}(\text{x}_\text{i} + \delta), t)$



Distribution loss:
Adversarial and target samples should follow a similar distribution.
$\text{MMD}_\text{F}(\text{X}_\text{t}, \text{X}_\text{i} + \Delta)$

Contribution: Designed SIA that effectively enforces statistical undetectability as a constraint.

Evaluating SIA against a range of anomaly detectors

Distributional detectors: Two different methods

- Defender needs to inspect **less than 50 samples** to detect prior attacks.
- **Over 1000 samples** to detect adversarial examples crafted by SIA.

Individual detectors: Five published methods that inspect each sample individually.

- Close to **random chance** detection performance against SIA.
- No customization for specific detectors.

Contribution: SIA is a general-purpose adaptive attack against a range of detectors.

Find us at the poster session for more details!

Hall E #323 - Wed 20 Jul 6:30pm

THANK YOU FOR LISTENING!

cankaya@umiacs.umd.edu