Automatic Discovery of Adaptive Attacks on Adversarial Defenses
Chengyuan Yao, Pavol Bielik, Petar Tsankov, Martin Vechev

Introduction

Adversarial defenses are proposed to address the problem of adversarial examples, but the authors of many defenses provide over-estimated robustness evaluations. These defenses are broken later with handcrafted adaptive attacks which are designed to reflect the defense mechanism, yet this approach requires strong domain expertise.

Our Work: We present an extensible tool $A^3$ that defines a search space over reusable blocks and automatically discovers an effective attack given the defense.

Motivation

Example Defenses

Robustness by authors

Handcrafted attacks (Tramer et al. 2020)

53% 15%

57% 5%

51% 0.2%

Our work: automate this adaptive process

Robustness Evaluation Paradigms

(a) Handcrafted adaptive attacks (Tramer et al. 2020)

(b) Ensemble of fixed attacks (Croce and Hein 2020)

(c) Adaptive attack search (Our Work)

Network Transformation


Loss Functions

- Difference between targeted and untargeted loss is the sign of the loss function.
- Logits/Probs means whether to add a softmax to logits.

Results

- $A^3$ is evaluated on 23 diverse defenses.
- Compared with AutoAttack (AA), the state of art ensemble of fixed attacks (Croce and Hein 2020).
- 10 cases: 3.0%-50.8% additional adversarial examples.
- 12 cases: $\approx 2x$ faster attack time. AutoAttack contains expensive but ineffective attacks.

Network Search

Goal: Find the best surrogate model $t$ to apply attack with. We use to generate adversarial images but test to evaluate.

Search: Exhaustive search. Use PGD as the test attack to evaluate each candidate.

Complexity: Cheap to perform

Overview of $A^3$