Defending the First-Order: Using Reluplex to Verify the Adversarial Robustness of Neural Networks to White Box Attacks

Adam Pahlavan (adampah@stanford.edu), Daniel Lee (dan9lee@stanford.edu), Justin Rose (justrose@stanford.edu)

Motivation

- **Adversarial Attacks: Small, imperceptible changes** to an image can easily **fool neural networks**
  - Security concern in safety-critical applications such as autonomous driving
  - Difficult to provide performance guarantees for models susceptible to attack
- **Adversarial Defenses:** Various published approaches for white-box-secure defenses
  - Evaluating defenses against **first-order gradient-based attacks** is the de-facto benchmark[1]
    - Projected gradient descent (PGD) and Fast-Gradient Sign Method (FGSM)

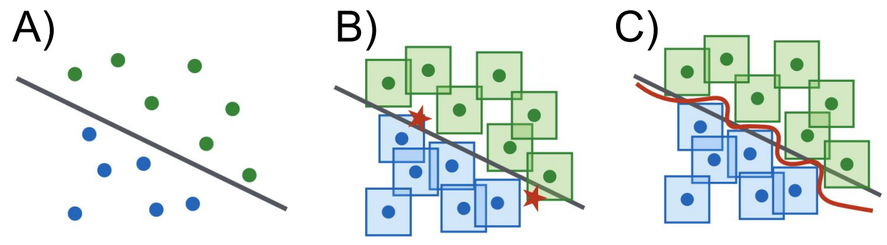

Fig. 1. (A) Simple linear classifier. (B) Classifier is susceptible to adversarial examples off the data manifold. (C) Adversarially trained classifiers learn robustness to nearby adversarial examples. Adapted from Madry[1].

Problem Statement

Are first-order attacks a good benchmark for verifying the adversarial robustness of a neural network? Do first-order defenses generalize to non-first order attacks?

Limitations of Current Science and Approach

- Neural networks are non-linear and nonconvex, and verifying even simple properties about them (such as finding the closest adversary) is NP-hard[2]
  - As a result, state-of-the-art adversarial defenses are constrained to benchmarking robustness against first-order gradient-based attacks
- We overcome this limitation with **Reluplex** (developed by Katz et. al[2]), a tool to verify the satisfiability of neural networks given input and output constraints
  - Reluplex is **sound** and **complete**: given a set of input and output constraints, it will never miss a satisfying condition
  - Due to Reluplex's current difficulties in scaling to larger networks, we study a multi-layer perceptron with one hidden layer with 50 neurons

Research Goal 1: Do first-order methods well approximate the closest adversary?

First-Order vs. Reluplex Attacks

Vanilla Model Accuracy: 97% test, 14% adversaries

- **Fast-Gradient Sign Method:** First-order iterative-optimization attack to find norm-bounded adversary
- If the adversary only has first-order information about a network, FGSM well-approximate the closest adversary[3]
- **Whether non-first-order methods can generate closer attacks is an open research question**

$$x^{t+1} = x^t + \epsilon * \text{sgn}(\nabla_x L(\theta, x, y))$$
$$\text{subject to } |x^{t+1}|_\infty \leq \delta$$

Equation 1. FGSM update rule

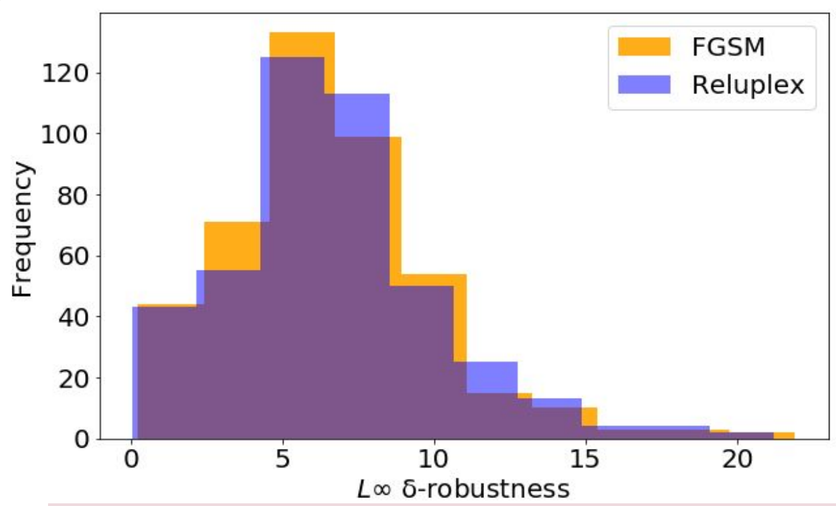

Fig. 2. Similar δ-robustness indicates Reluplex's attacks are not significantly closer than FGSM's

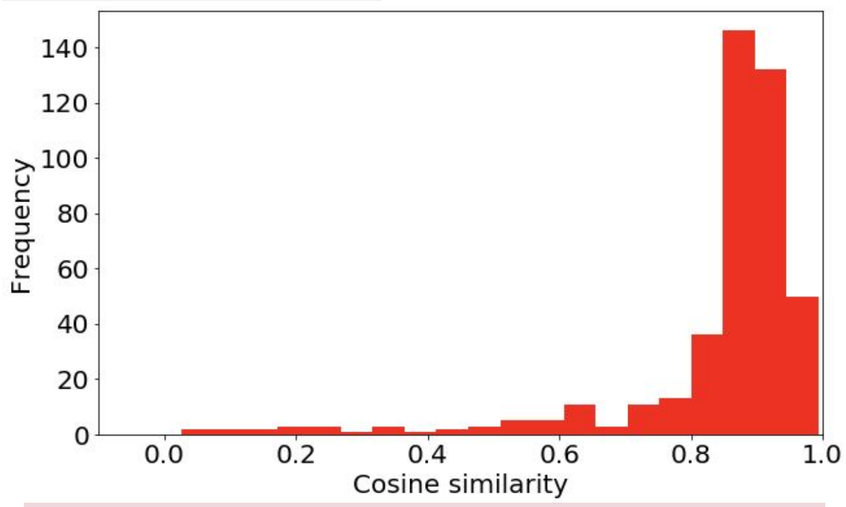

Fig. 3. High cosine similarity indicates Reluplex's attacks are in a similar direction to FGSM's

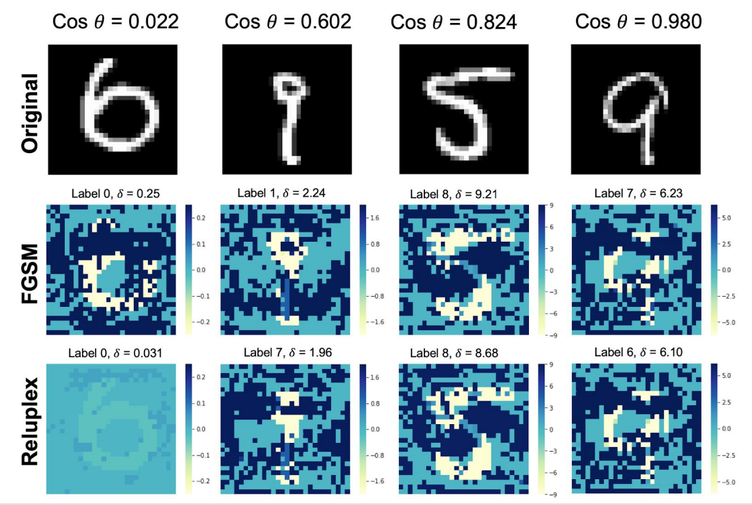

Fig. 4. Perturbations are visually similar in both methods, indicating Reluplex finds similar attacks

Research Goal 2: Do state-of-the-art adversarial defenses generalize to non-first-order attacks?

- Several white-box adversarial defenses have been shown to increase robustness against first order attacks
  - Research suggests many state-of-the-art techniques are **shallow** and only provide defenses to first-order attacks by **obfuscating first-order gradients**[4]

Case Study: Adversarial Logit Pairing

Robust Model Accuracy: 96.5% test, 89% adversaries

- Matches logits from images and their corresponding FGSM-generated adversaries by minimizing the loss

$$J(\boldsymbol{X}, \boldsymbol{\theta}) + \lambda \frac{1}{m} \sum_{i=1}^{m} L(f(\boldsymbol{x_i}, \boldsymbol{\theta}), f(\widetilde{\boldsymbol{x}}_i, \boldsymbol{\theta}))$$

Equation 2. Adversarial logit pairing loss function

for clean images $x_i \in \{x_1, x_2, ..., x_m\}$
adverserial images $\widetilde{x}_i \in \{\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_m\}$
cross entropy classifier loss $J(\boldsymbol{X}, \boldsymbol{\theta})$

- **Baseline:** Adversarial logit pairing significantly improves robustness to first-order attacks (Fig. 5)

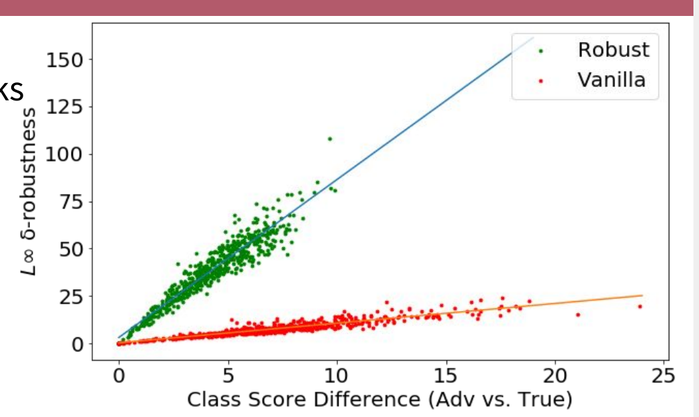

Fig. 5. Baseline robustness comparison for FGSM-generated adversaries for vanilla (R^2=0.93) and robust network (R^2=0.95).

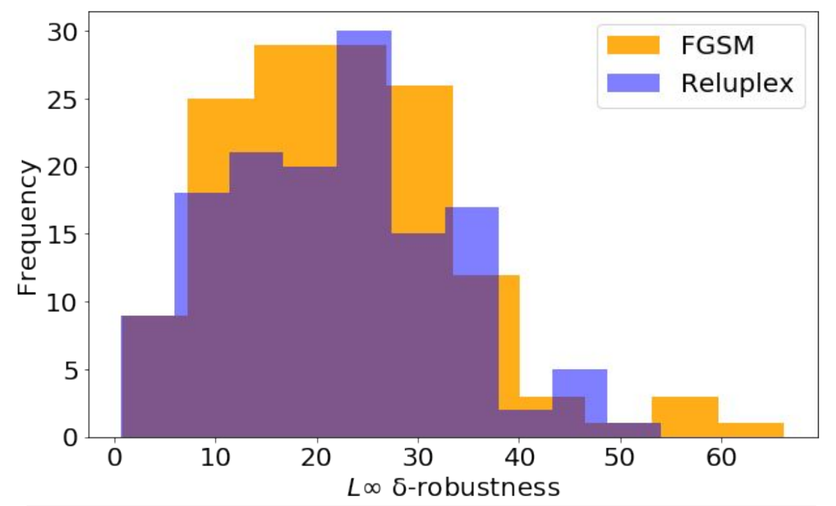

Fig. 6. Similar δ-robustness indicates Reluplex's attacks are not significantly closer than FGSM's

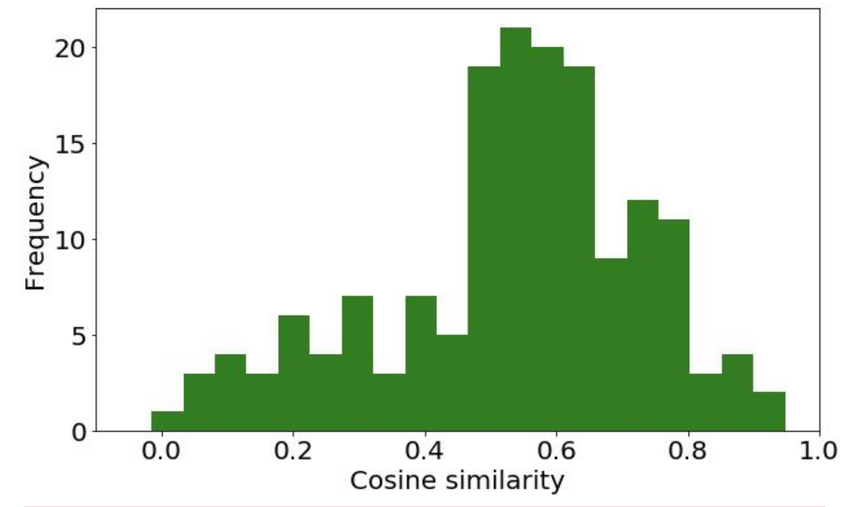

Fig. 7. Lower cosine similarity indicates Reluplex's attacks explore different directions than FGSM's

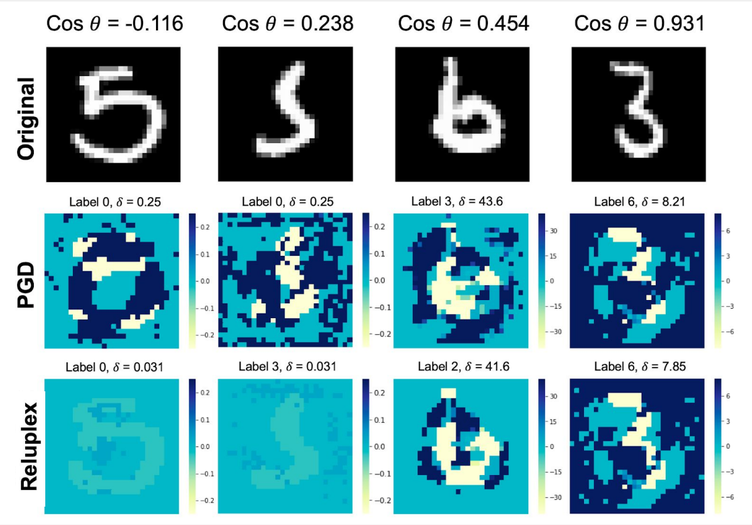

Fig. 8. Perturbations are dissimilar: FGSM attacks on robust model more sensitive to initialization

Conclusion

I. First-order attacks provide a close approximations of the closest adversary over the entire input domain. **Supports first-order attacks as good benchmarks for evaluating robustness**

II. Adversarial training against first order attacks generalizes to all attacks. **Supports that first order defenses are good universal defenses**

III. We conjecture first-order methods well-approximate the closest adversary because our network can only behave in a **limited range of linear modes** since **ReLUs are fixed in a local region**

IV. Future work can expand the current study to large and deep networks, where **non-first-order attacks** can exploit the **more complicated loss surface** and **greater variation in linear modes**

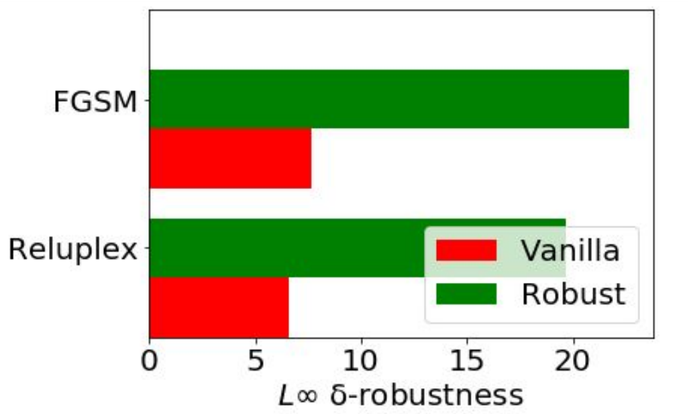

Fig. 9. Average δ-robustness comparison for the vanilla and robust networks. Adversarial training significantly improves robustness under FGSM and Reluplex attacks. Reluplex finds slightly closer adversaries than FGSM.