Certify or Predict: Boosting Certified Robustness with Compositional Architectures

Mark Müller
Mislav Balunovic
Martin Vechev
Adversarial Examples

\[ x + .007 \times \text{sign}(\nabla_x J(\theta, x, y)) = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

- `x`: "panda" 57.7% confidence
- \( \text{sign}(\nabla_x J(\theta, x, y)) \): "nematode" 8.2% confidence
- \( x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \): "gibbon" 99.3% confidence

Neural Network Verification

- Robustness property:

\[ \arg\max_i h(x)_i = \arg\max_i h(x')_i \quad \forall x' \in B^\infty(x) \]
Problem Statement

- Adversarial accuracy requires increased network capacity
- Verification gets increasingly difficult with network depth
  ➔ Small, provably trained networks have low standard accuracy
  ➔ ACE: Compose networks with different strengths
ACE – Compositional Architecture

• For every sample decide whether to use core- or certification-network

• Key components:
  • Deep standard network
  • Shallow provable network
  • Selection mechanism
    • Train network to predict certification difficulty
    • Evaluate certification network entropy
Effectiveness of Selection

- Strong separation of samples based on certifiability
- Significantly increased accuracy of the certification-network on the selected sample subset
ACE Results

- Significant reduction in certified accuracy loss, for gains in natural accuracy

- Effect observed across:
  - Network architectures
  - Perturbation sizes
  - Datasets
  - Certification and training methods

\[
\text{CIFAR-10 } \varepsilon_\infty = \frac{2}{255}
\]

\[
\text{CIFAR-10 } \varepsilon_\infty = \frac{8}{255}
\]

\[
\text{TinyImageNet } \varepsilon_\infty = \frac{1}{255}
\]

Balunovic, Mislav, and Martin Vechev. “Adversarial training and provable defenses: Bridging the gap.” ICLR 2019
Thank you for your attention!

Paper and Code:
https://www.sri.inf.ethz.ch/publications/mueller2021boosting

Poster Session 10