Defenses Against Adversarial Examples

Neil Gong

Certifiably robust classifier

- A classifier is (p, ε)-certifiably robust for x, if no adversarial perturbation whose L_p norm is no larger than ε exists.

- Verification
  - Given a classifier and x, verify whether the classifier is (p, ε)-certifiably robust for x

- Certification
  - Given a classifier and x, deriving p and ε

Verification via interval analysis

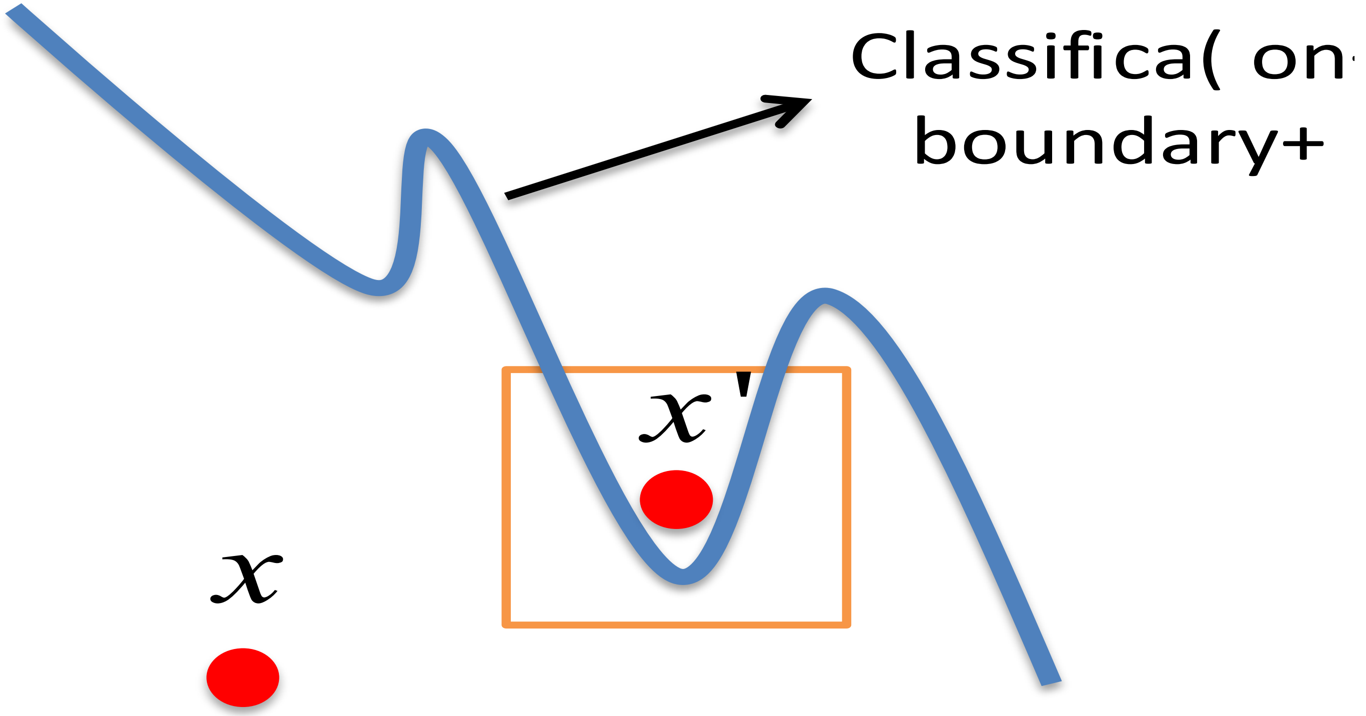
- Given x, p=∞, ε, we propagate the intervals from the input to the output

- Limitations
  - False negatives
  - Limited to p=∞
  - Not effective for certain classifiers

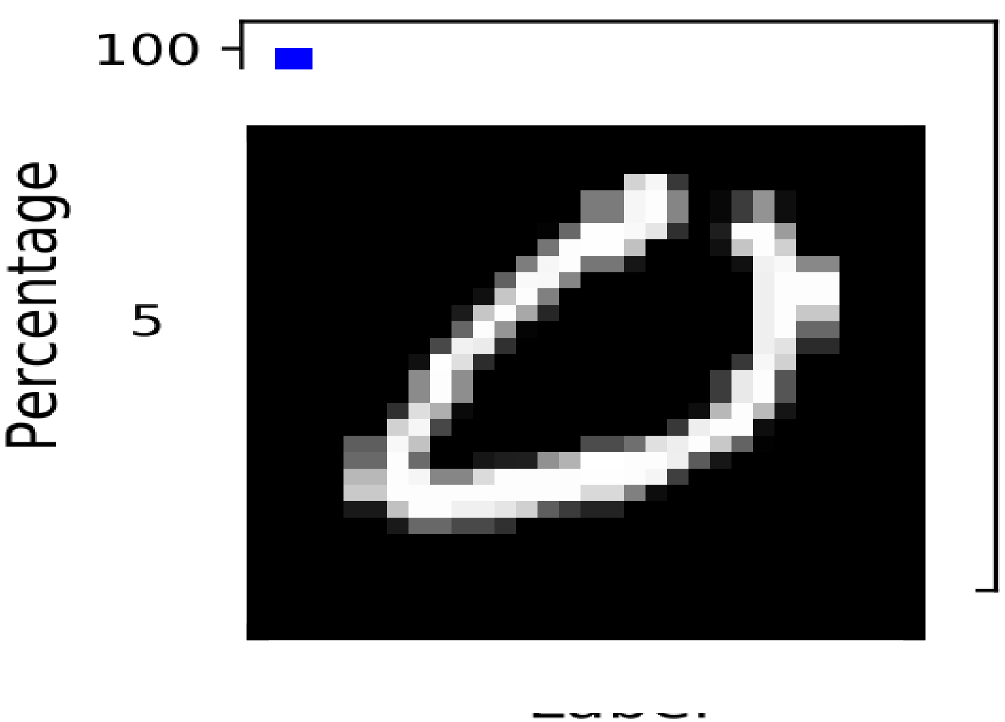
Certification via randomized smoothing

- Given a classifier and x, deriving p and ε
- Many methods have been developed
- Randomized smoothing
  - Applicable to any classifier
  - Scalable to large neural networks

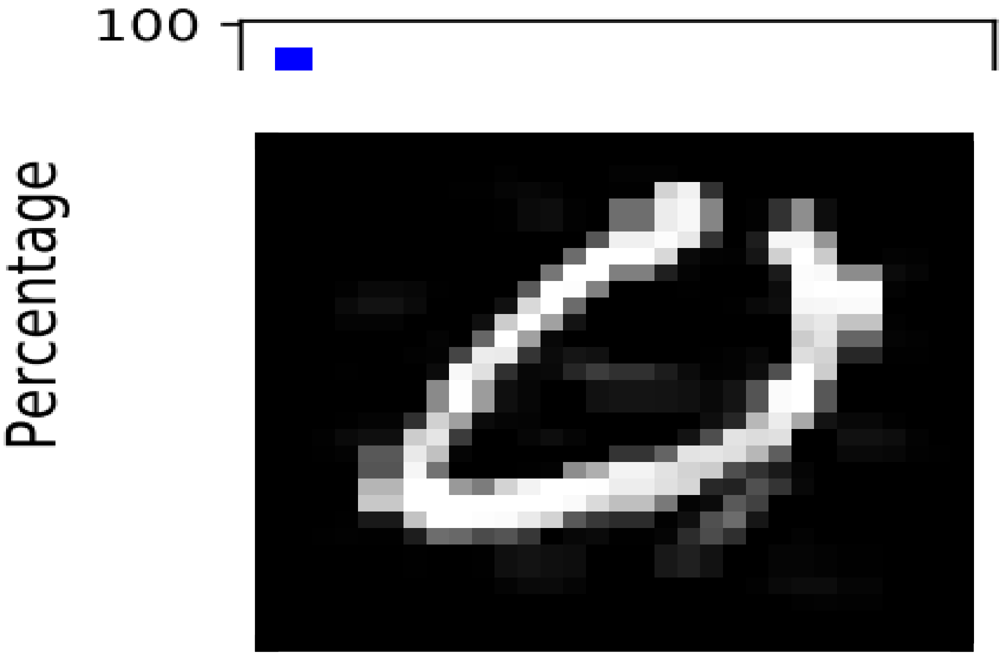
Adversarial example is close to classification boundary?



Measuring Adversarial Examples

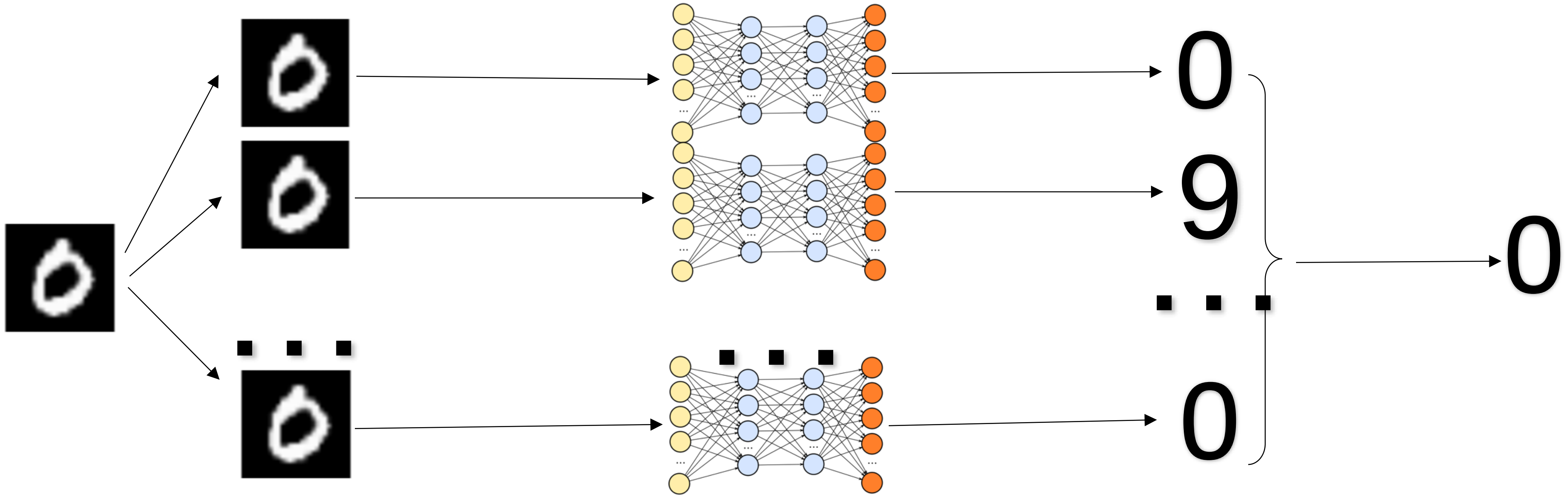


A normal example: digit 0



An adversarial example
with a target label 9

Randomized smoothing



Formal definition of randomized smoothing

- Input
  - a classifier f
  - an example x
  - a noise distribution
- Output
  - $\mathrm{g}(x) = \underset{c}{\mathrm{argmax}} \Pr(f(x + r) = c)$

Deriving (p, ε)

- Noise is isotropic Gaussian distribution

- $\mathrm{g}(x + \delta) = C_A$ when $|\delta|_2 \leq \mathcal{E}$

$$\mathcal{E} = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right)$$

Certified radius

Tightness of the bound

- Given
  - No assumptions on the classifier f
  - Randomized smoothing with Gaussian noise
- The derived bound is tight

Estimating the label probabilities

- Sampling a large number of noise
- Predicting labels for the noisy examples
- Estimating label probabilities with probabilistic guarantees

Generalization to top-k

- Input
  - a classifier f
  - an example x
  - a noise distribution
- Output
  - $p_c = \Pr(f(x + r) = c)$
  - The smoothed classifier predicts k labels with the largest label probabilities

- A label is among the top-k labels if the adversarial perturbation is bounded

Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing

Training to improve certified accuracy

- Adding random noise during training
- Adding certified radius as a regularization term

$$\underbrace{\mathbf{1}_{\{g_\theta(x)\neq y\}}}_{\text{0/1 Classification Error}} + \underbrace{\mathbf{1}_{\{g_\theta(x)=y, CR(g_\theta;x,y)<\epsilon\}}}_{\text{0/1 Robustness Error}}$$

MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius

Randomized smoothing

- Strengths
  - Applicable to any classifier
  - Scalable to large classifier

- Limitations
  - Efficiency – need many predictions
  - Probabilistic guarantees