Defenses Against Adversarial Examples

Neil Gong

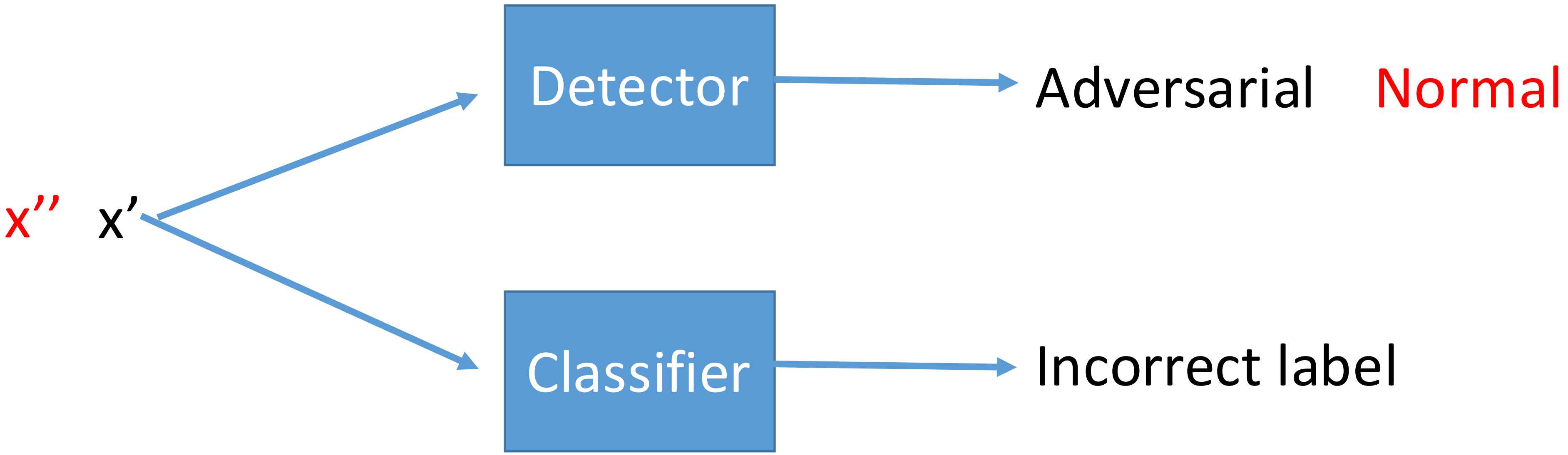
Defending against adversarial examples

- General philosophy for security solutions
  - Prevention
  - Detection
  - Response
- Prevention
  - robust classifiers
- Detection
  - detecting adversarial examples
- Response
  - manual labeling?
  - collecting more data?

Detecting adversarial examples

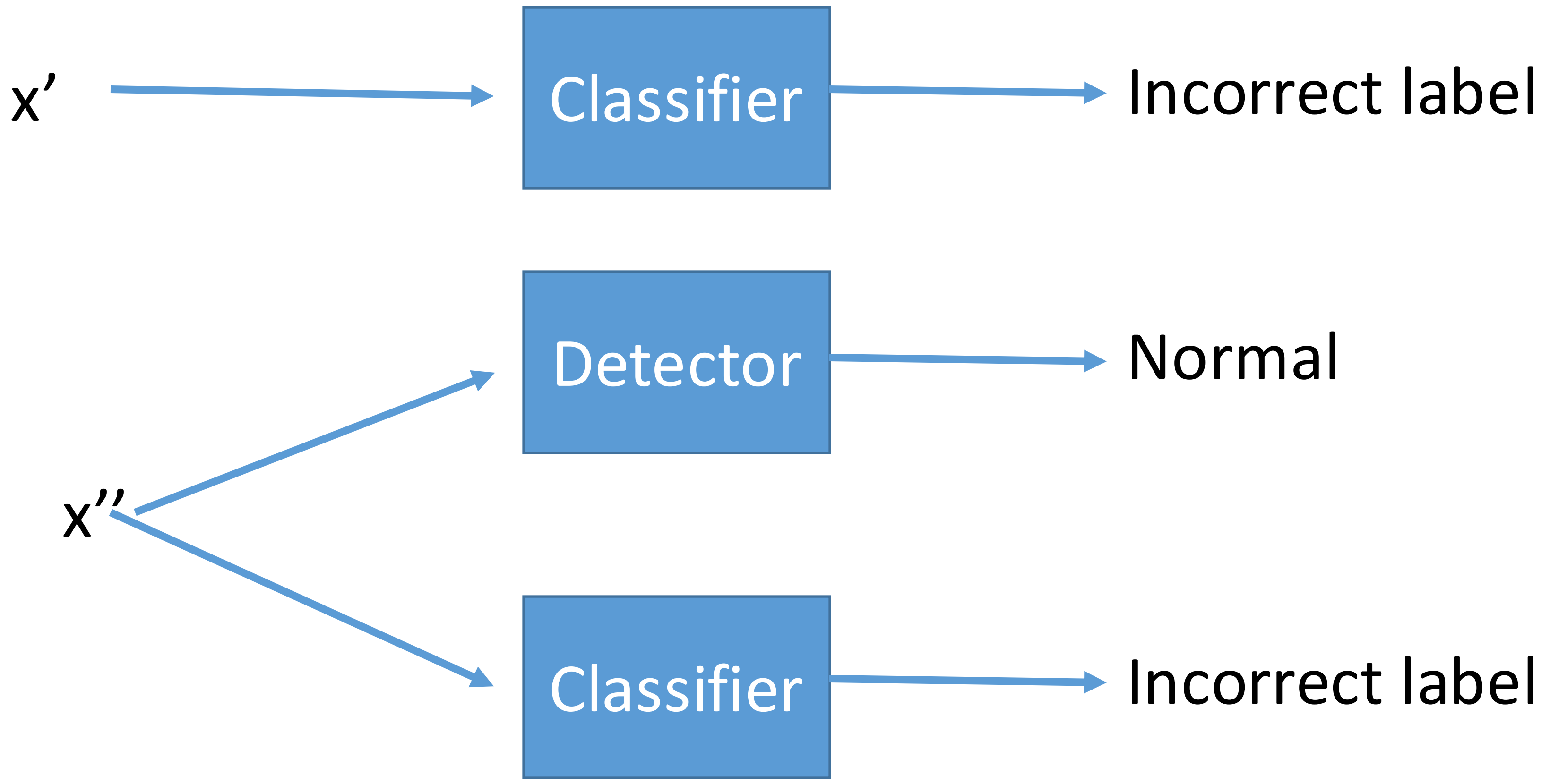
- Binary classification
  - Normal example vs. adversarial example
- Add one more label "adversarial"
  - E.g., 0, 1, 2, …, 9, adversarial
- Extracting features and building detectors

Challenges of detecting adversarial examples



Attackers are adaptive

Evaluating a detection method



Evaluating a detection method

- Metric 1
  - Whether human perceives x'' and x as the same
  - no-> Detection is effective
  - Hard to implement

- Metric 2
  - $d(x',x)$ vs. $d(x'', x)$
  - $d(x'', x) > d(x',x)$ -> detection is effective
  - $d(x'', x) - d(x',x)$ measures effectiveness
  - Consider strong adaptive attacks

Response

- Manual labeling
- Collecting more data
  - Other sensor data
- Forensics
  - Root cause analysis
  - Attack source
- Recovery

Prevention – robust classifiers

- Empirically robust classifier
  - A particular attack cannot find adversarial example within a L_p norm ball
  - (p, ε)-robust against an attack for x, if the attack does not find adversarial perturbation whose L_p norm is no larger than ε.

- Certifiably robust classifier
  - No adversarial examples exist within a L_p norm ball.
  - (p, ε)-certifiably robust for x, if no adversarial perturbation whose L_p norm is no larger than ε exists.

Training empirically robust classifier

An attack $$\max_{\delta \in B_p(x,\varepsilon)} L(x+\delta, y|\theta)$$

Adversarial training $$\min_{\theta} \sum_{(x,y)} \max_{\delta \in B_p(x,\varepsilon)} L(x+\delta, y|\theta)$$

Adversarial training

$$\min_{\theta} \sum_{(x,y)} \max_{\delta \in B_p(x,\varepsilon)} L(x+\delta, y|\theta)$$

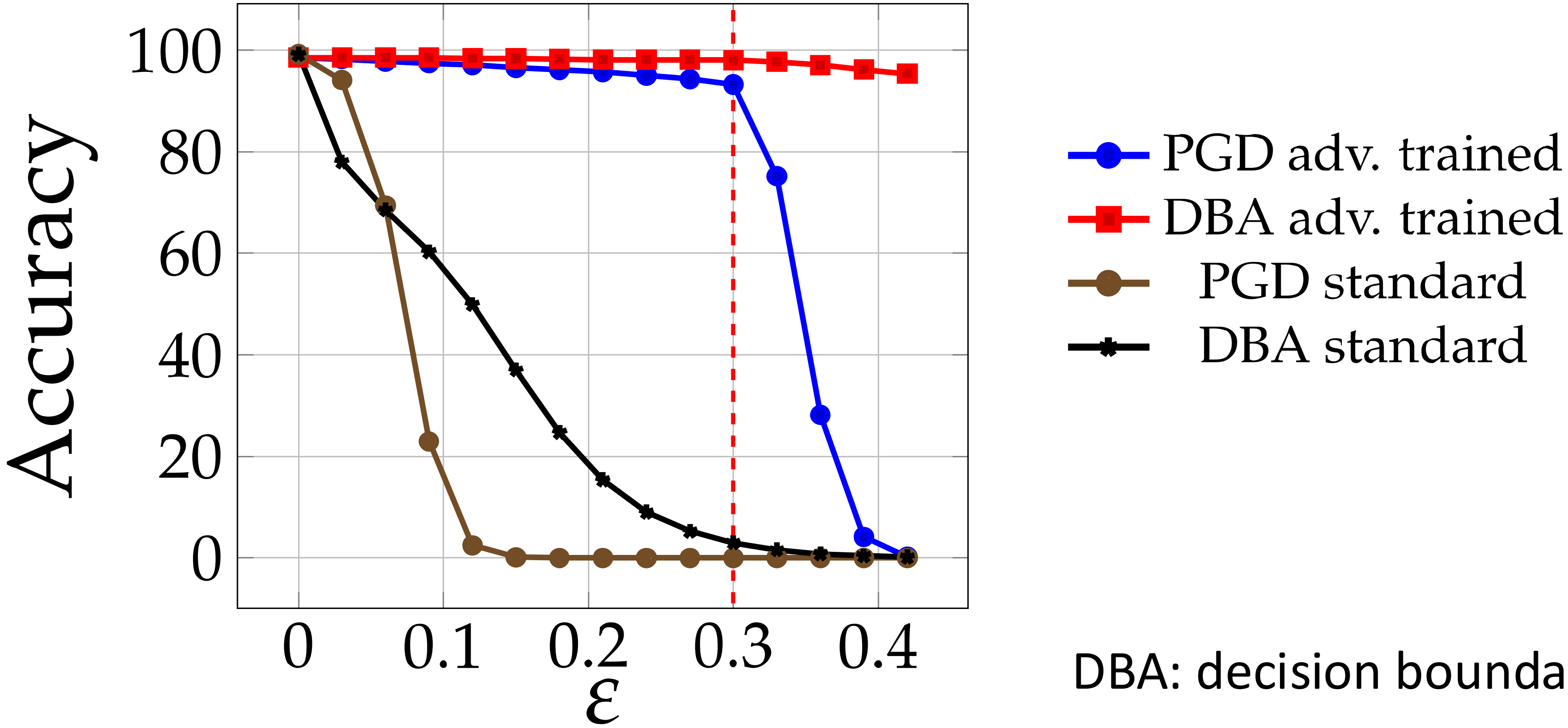
- Alternate between max and min
- Inner max
  - Finding adversarial perturbation δ, e.g., Projected Gradient Descent (PGD)
- Outer min
  - Updating model parameters θ using both normal and adversarial examples

Issues of adversarial training

- No certifiable guarantee

- May not be empirically robust against unseen attacks
    - Use multiple attacks during training

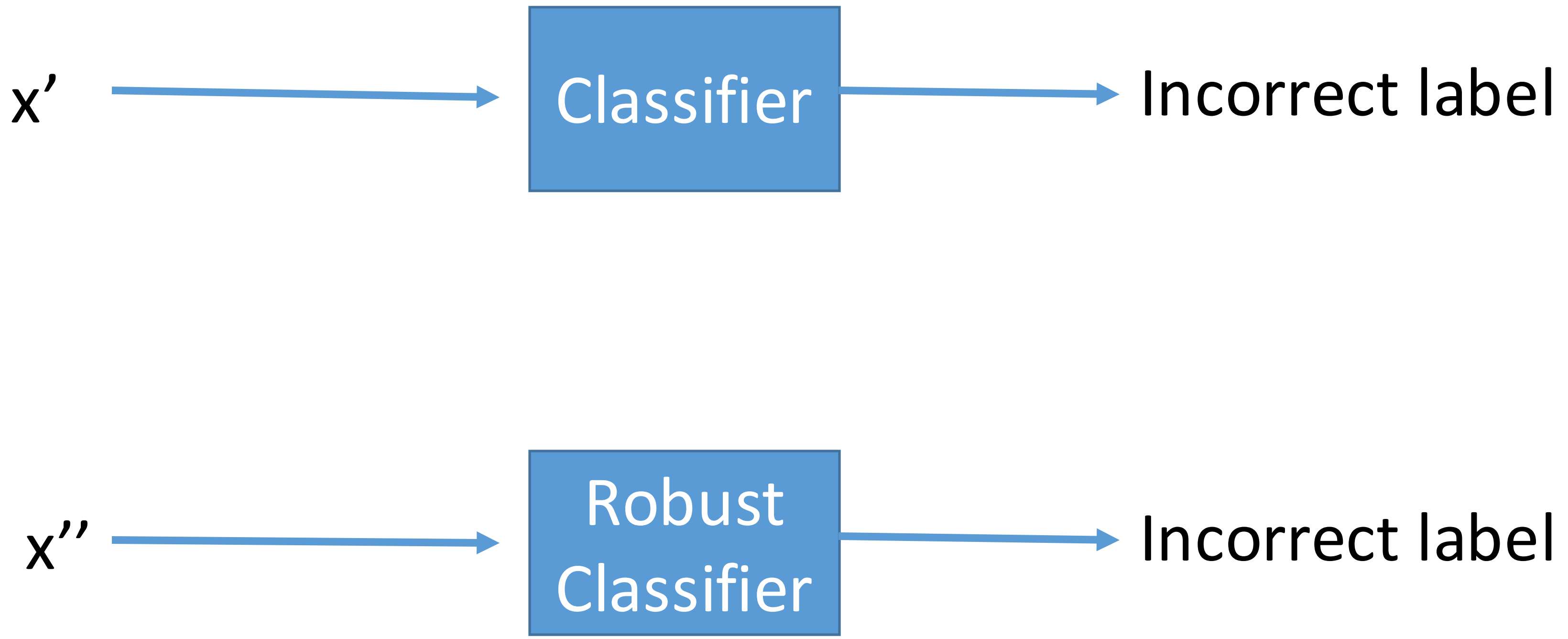
- May not be robust to perturbation larger than ε used in training



DBA: decision boundary attack

(a) MNIST, ℓ_∞-norm

Evaluating an empirically robust classifier



Evaluating an empirically robust classifier

- Metric 1
  - Whether human perceives x'' and x as the same
  - no-> defense is effective
  - Hard to implement

- Metric 2
  - $d(x',x)$ vs. $d(x'', x)$
  - $d(x'', x) > d(x',x)$ -> defense is effective
  - $d(x'', x) - d(x',x)$ measures effectiveness
  - Consider strong adaptive attacks