Notice

- Due dates
  - Homework 1 (01.17)
  - Written paper critiques (on 01.19)
  - Term Project (Sign-up by 01.19) **[Want Random by 01.17?]**
- Sign-up (on Canvas)
  - Scribe Lecture Note
  - In-class Paper Presentation / Discussion
- Zoom link for the class
  - Please email me if you have (to be quarantined, illness, …)

CS 499/599: Machine Learning Security

01.12: Adversarial Examples (AE) 3

Mon/Wed 12:00 – 1:50 pm

Instructor: **Sanghyun Hong**

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab

Recap

- ML Matters
- Evasion (Test-time Adversarial) Attack
  - Threat Model
  - Attack:
    - FGSM Attack
  - Mitigation:
    - Adversarial Training (AT)

Topics for Today

- AE←Security
  - Practical considerations
  - Iterative Methods
  - Real-world exploitation
- AE←Security
  - Motivation
  - C&W Attack
  - Conclusions (and Implications)
- AE←ML
  - Motivation
  - PGD Attack
  - Conclusions (and Implications)

Alexey et al., Adversarial Examples in the Physical World

Motivation

- Remaining Questions:
  - RQ3: Can an adversary **exploit adversarial examples in practice**?

Motivation – cont'd

- AE in the numerical world ≠ AE in the physical world
  - Numerical perturbations by FGSM lead to the input values like 34.487
  - In the pixel space, such perturbations do not exist (*i.e.*, quantized pixel values)
  - One may take only classification results with a high probability (*e.g.*, > 0.8)
  - Many others…

- An example (CIFAR-10)
  - Craft AEs on a DNN model (~an error rate of 99.9%)
  - Store these AEs into PNG files
  - Upload them to object recognition services (~an error rate of 10%)

Revisit the FGSM Method

- Given
  - A test-time input (X, y); each pixel in $X \sim$ [0, 255]
  - A NN model f and its parameters θ
  - A loss (or a cost) function $J(X, y)$
- Find
  - An AE X^{adv} such that $f\left(X^{adv}\right) \neq y$ and $||X^{adv} - X||_{\infty} \leq \varepsilon$

$$\boxed{\boldsymbol{X}^{adv} = \boldsymbol{X} + \epsilon\, \mathrm{sign}\big(\nabla_X J(\boldsymbol{X}, y_{true})\big)}$$

Basic Iterative Method

- Objectives
  - To **scale** numerically small perturbations (*i.e.*, pixel values ~ [0, 255])
  - To craft **powerful** AEs

- BIM Method
  - Run FGSM over multiple iterations

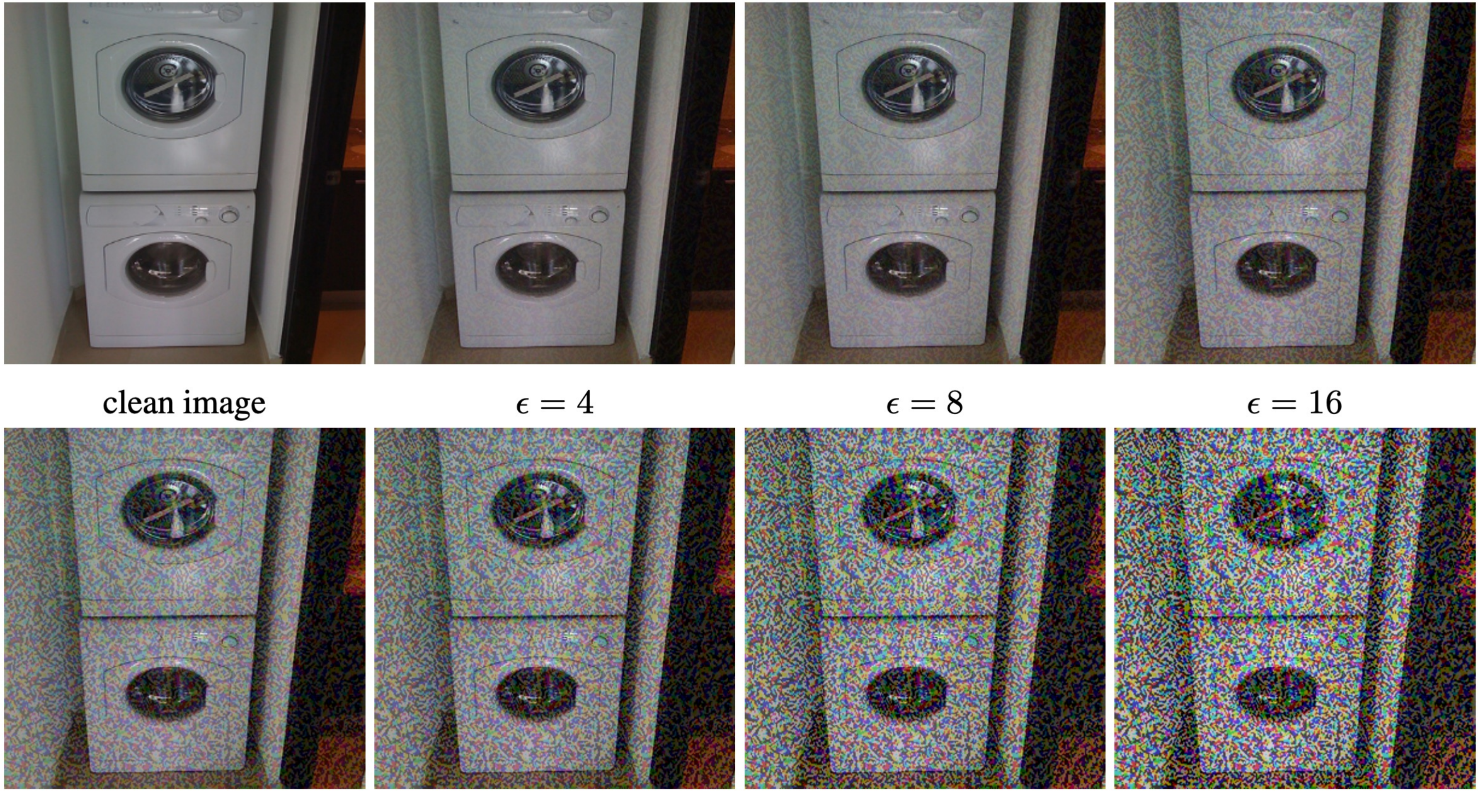
$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\Big\{\boldsymbol{X}_N^{adv} + \alpha\,\mathrm{sign}\big(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\big)\Big\}$$

- Iterative Least-Likely (ILL) Class Method
  - Choose a desired class as the class with the lowest logit value (y_{LL})

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} - \alpha\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{LL})\right)\right\}$$

Empirical Results on the ImageNet Inception-v3

• Co



clean image · $\epsilon = 4$ · $\epsilon = 8$ · $\epsilon = 16$

$\epsilon = 24$ · $\epsilon = 32$ · $\epsilon = 48$ · $\epsilon = 64$

–

–

–

–

RQ 3: Real-World Exploitation

- Setup
  1. Craft AEs, store them in PNG, and print them
  2. Take photos of printed AEs with a cell phone
  3. Resize and center-crop the images from 2
  4. Run classification on the images from 3

(a) Printout (b) Photo of printout (c) Cropped image

RQ 3: Real-World Exploitation

- Observations
  - AEs work in the physical world
    - Misclassification rate is higher in AEs than what we observe with clean examples
    - Chances increase when we increase the perturbations (*i.e.*, eps from 2 to 16)
  - Prefiltering can reduce the misclassification significantly
    - **Prefilter:** only accept the classification with a high probability > 0.8
    - It reduces an error rate by 40 – 90%
  - Can we think some other system-level defenses?

RQ 3: Still, I Can't Believe It Works

- Link, Link, Link

Conclusions

- Lessons
  - RQ2: How can we find the adversarial examples **efficiently**?
    - BIM (Basic Iterative method)
    - ILL (Iterative Least-Likely class method)
  - RQ3: Can an adversary **exploit adversarial examples in practice**?
    - **Highly Likely!**

Topics for Today

- AE←Security
  - Practical considerations
  - Iterative Methods
  - Real-world exploitation
- AE←Security
  - Motivation
  - C&W Attack
  - Conclusions (and Implications)
- AE←ML
  - Motivation
  - PGD Attack
  - Conclusions (and Implications)

Carlini et al, Towards Evaluating the Robustness of Neural Networks

Explosive Interests in AEs

- Many Attacks
  - FGSM
  - BIM (ILL-Class)
  - JSMA
  - DeepFool
  - …

- Defense Proposals
  - NN architectures resilient to AEs
  - Adversarial Training [?!]
  - Defensive Distillation

Motivation

- Research Questions:
  - RQ 1: **What attacks** should we choose for evaluating NN robustness?
  - RQ 2: How much are the existing defenses **effective against AEs**?

Revisit the Threat Model

- Given
  - A test-time input (x, y); each element in $x \sim$ **[0, 1]**
  - A NN model f and its parameters θ
- Goal
  - Find an x^{adv} such that $f(x^{adv}) \neq y$ while $||x^{adv} - x||_p \leq \varepsilon$

Revisit the Threat Model – cont'd

- Given
  - A test-time input (x, y); each element in $x \sim$ **[0, 1]**
  - A NN model f and its parameters θ
- Goal
  - Find an x^{adv} such that $f(x^{adv}) = \boldsymbol{t}\ (\boldsymbol{t} \neq \boldsymbol{y})$ while $||x^{adv} - x||_p \leq \varepsilon$
- Three scenarios (depends on how we choose $\boldsymbol{y^t} = f(x^{adv})$)
  - Best-case
  - Average-case
  - Worst-case

Revisit the Threat Model – cont'd

- Given
  - A test-time input (x, y); each element in $x \sim$ **[0, 1]**
  - A NN model f and its parameters θ
- Goal
  - Find an x^{adv} such that $f\left(x^{adv}\right) = t$ $(t \neq y)$ while $||x^{adv} - x||_p \leq \varepsilon$
- Three scenarios (depends on how we choose $\boldsymbol{y^t} = f\left(x^{adv}\right)$)
  - Best-case
  - Average-case
  - Worst-case
- Perturbations
  - L_0, L_1, L_2, L_∞

RQ 1: How to Evaluate the Robustness of NNs?

- The Problem:

$$\begin{array}{ll} \text{minimize} & \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ \text{such that} & x+\delta \in [0,1]^n \end{array}$$

  - Somewhat computationally tractable problem
  - c: a hyper-parameter found by binary search

- Many Possible C (or f)
  - Refer to the paper ($f_1 \sim f_7$)

- Optimization: PGD, Clipped GD, Change of Variables (Refer to the paper)

RQ 1: How to Evaluate the Robustness of NNs?

- Carlini & Wagner (C&W) Attack:
  - L_2 Attack:

$$\text{minimize} \quad \|\frac{1}{2}(\tanh(w)+1) - x\|_2^2 + c \cdot f(\frac{1}{2}(\tanh(w)+1)$$

with f defined as

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa).$$

  - L_0 Attack: at each iteration, find pixel locations we don't want to perturb
  - L_∞ Attack: same as L_2

RQ 1: Evaluation Results

- Setup
  - MNIST, CIFAR-10, and ImageNet
  - on randomly chosen 1000 test-time images
- Baselines
  - FGSM, BIM, JSMA, and DeepFool
- Results:
  - C&W achieves 100% misclassification rate
    - It uses 2x – 10x less perturbations than the baselines
    - FGSM often shows 0 – 42% success rate (weak attack)

Motivation – revisit'ed

- Research Questions:
  - RQ 1: **What attacks** should we choose for evaluating NN robustness?
  - RQ 2: How much are the existing defenses **effective against AEs**?

Defensive Distillation

- The Key Idea
  - Increase the distillation temperature T so that classification becomes more confident

- Their Results
  - Reduces the misclassification by AEs
    - from 96% to 0% (MNIST)
    - from 88% to 5% (CIFAR-10)

RQ 2: How to Evaluate Defenses?

- Take-away
  - Use **strong (or the strongest) attacks** to evaluate defenses
  - Defense should also **break the transferability**

- Results:
  - C&W achieves 100% misclassification rate against defensive distillation
  - C&W's misclassification rate does not depend on the distillation temperature
  - When carefully crafted,
    - C&W AEs crafted on a model transfers to a model trained with defensive distillations
    - It transfer with 0 – 100% depending on the choice of k in [0, 40]

Conclusions

- Lessons
  - RQ 1: **What attacks** should we choose for evaluating NN robustness?
    - Not just existing attacks, but a strong baseline attack
  - RQ 2: How much are the existing defenses **effective against AEs**?
    - Defenses not evaluated with strong baseline attacks are weak
    - Defenses should break the transferability, too

Thank You!

Mon/Wed 12:00 – 1:50 pm

Instructor: **Sanghyun Hong**

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL
Secure AI Systems Lab