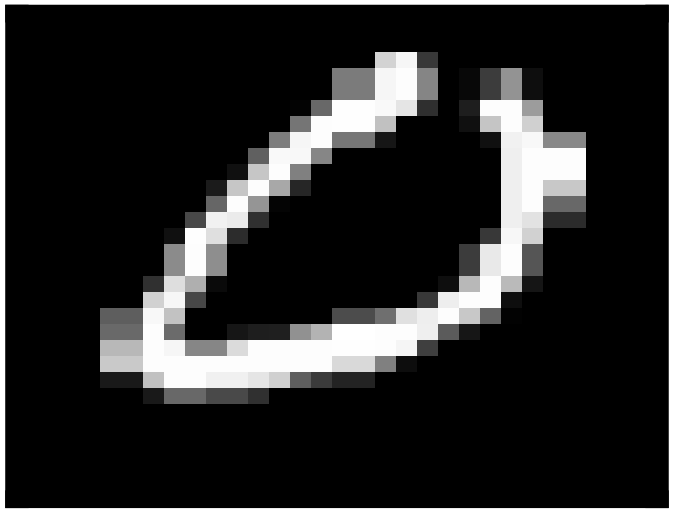
Adversarial Examples

Neil Gong

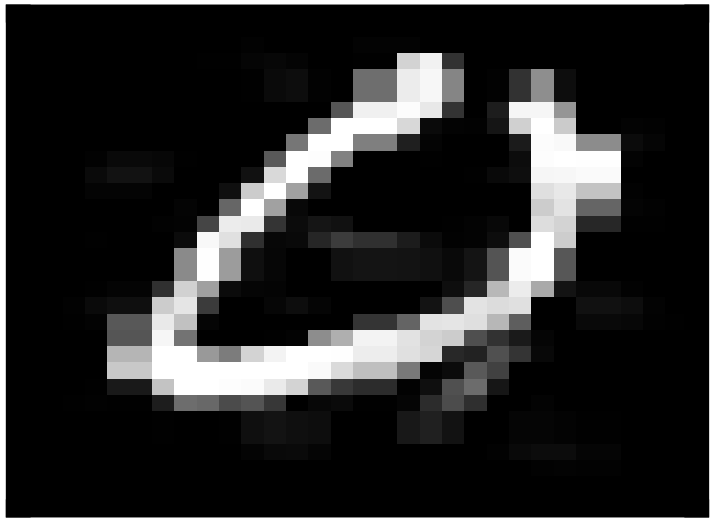
Today’s lecture

- What is adversarial example
- Why do we care
- How to find adversarial example

Adversarial Examples



Normal example: digit 0



Adversarial example:
predicted to be 9

Adversarial Examples

- Classifier *C*
- Normal example *x*
  - Image, text, audio, graph, software
- Perturb x to *x'*
  - Preserving semantics
- *C* misclassifies *x'*
  - *Targeted*: *C(x')=t,* an attacker-chosen target label
  - *Untargeted: C(x') ≠ C(x)*

Why do we care?

Stop sign to speed limit

Malware -> benign software

Spam -> non-spam

Privacy protection

Safety of generative AI

Guiding design of ML

Threat model

- Attacker’s goal
- Attacker’s background knowledge
- Attacker’s capability

Attacker’s Background Knowledge

No access
No box
Weakest attacker

Access to prediction API
Black box

Learning algorithm
Hyperparameter, e.g., neural network architecture
Training data

Model parameters
White box
Strongest attacker

How to Find Adversarial Examples - Image Domain

- Perturb x to *x'*
  - Preserving semantics
    - *Human perceives x' and x as the same*
    - *d(x,x') is small*

L_0, L_2, L_∞ norm of the noise x'-x

Minimize $d(x,x')$

Subject to (1) $C(x') = t$ or $C(x') \neq C(x)$

(2) x' is still an image

Solving the optimization problem

$$\begin{aligned} \text{minimize} \quad & \|\delta\|_p + c \cdot f(x+\delta) \\ \text{such that} \quad & x + \delta \in [0,1]^n \end{aligned}$$

x’

Box constraints

Loss function

$$f_1(x') = -\text{loss}_{F,t}(x') + 1$$

$$f_2(x') = (\max_{i \neq t}(F(x')_i) - F(x')_t)^+$$

$$f_3(x') = \text{softplus}(\max_{i \neq t}(F(x')_i) - F(x')_t) - \log(2)$$

$$f_4(x') = (0.5 - F(x')_t)^+$$

$$f_5(x') = -\log(2F(x')_t - 2)$$

$$f_6(x') = (\max_{i \neq t}(Z(x')_i) - Z(x')_t)^+$$

$$f_7(x') = \text{softplus}(\max_{i \neq t}(Z(x')_i) - Z(x')_t) - \log(2)$$

Box constraints

- Projected gradient descent

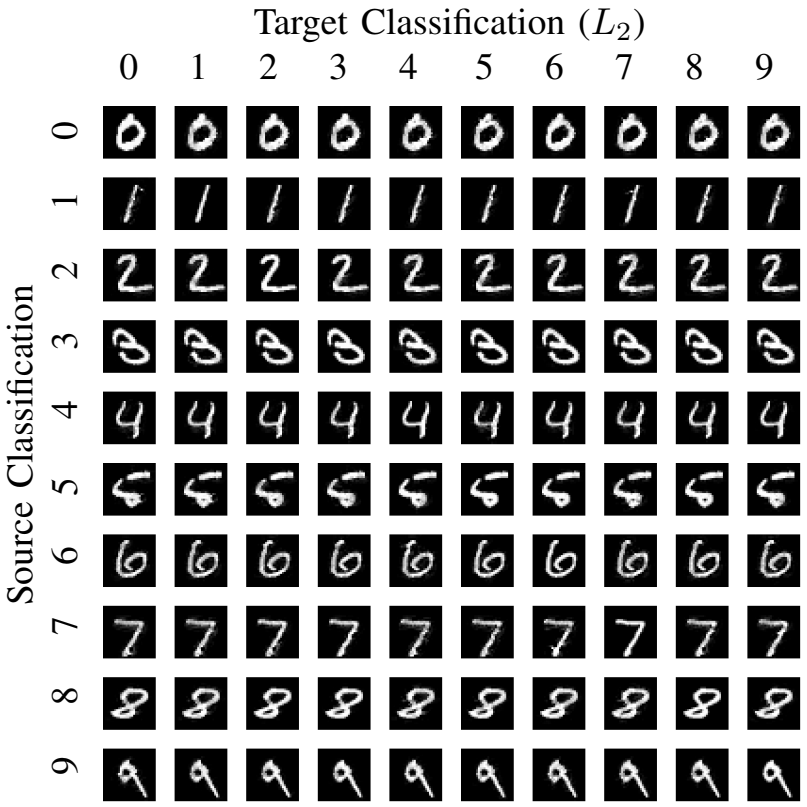
- Clipped gradient descent
  - Incorporate clipping into objective function

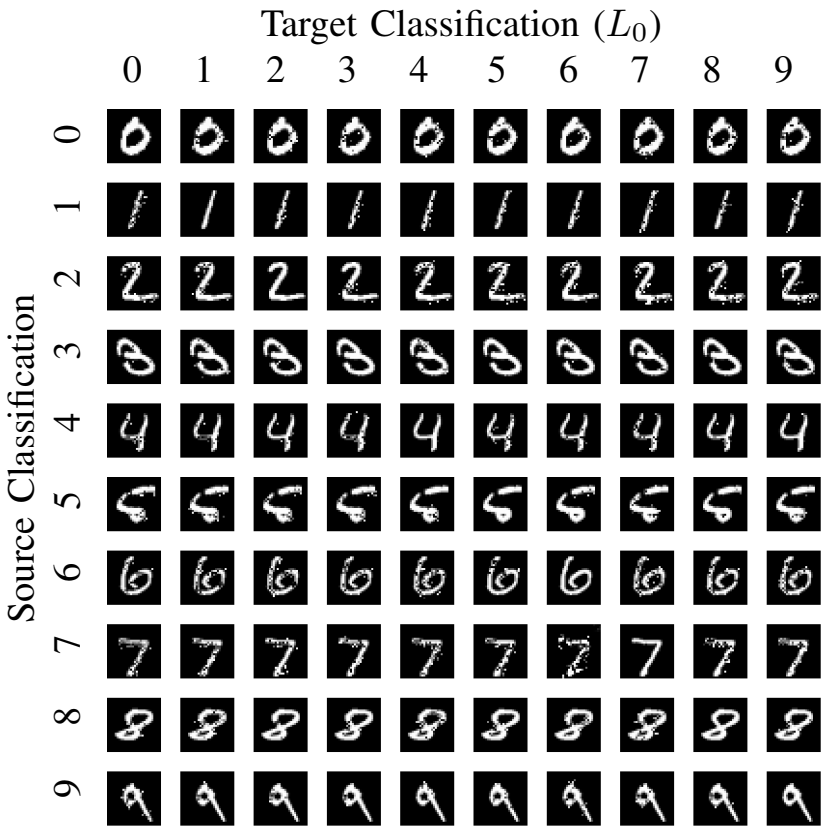
$$f(\min(\max(x+\delta, 0), 1))$$

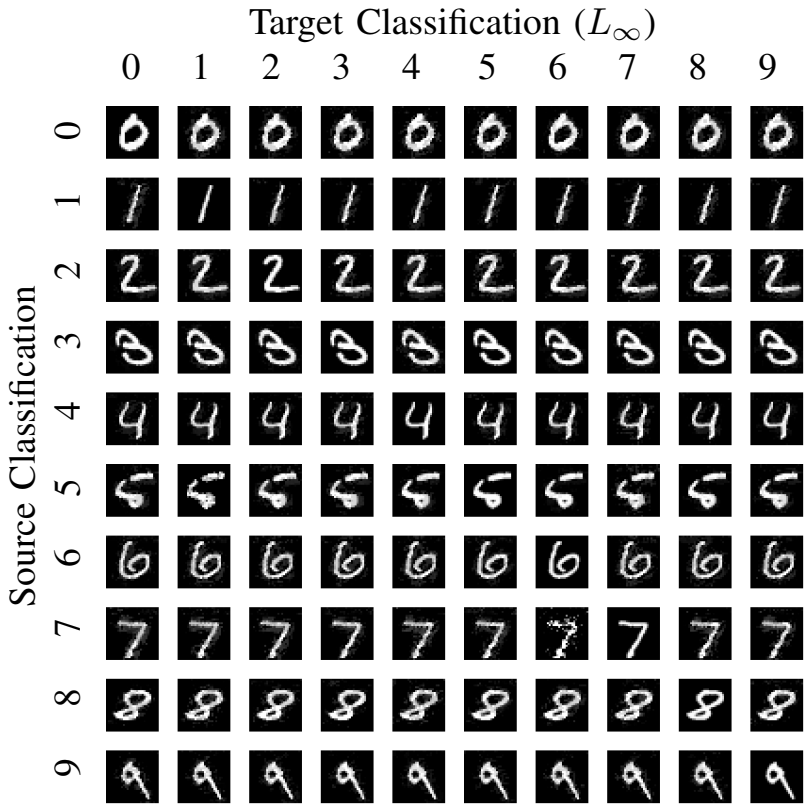
- Change of variables

$$\delta_i = \frac{1}{2}(\tanh(w_i) + 1) - x_i$$

Examples


Target Classification (L_2)
0 1 2 3 4 5 6 7 8 9
Source Classification
0 1 2 3 4 5 6 7 8 9


Target Classification (L_0)
0 1 2 3 4 5 6 7 8 9
Source Classification
0 1 2 3 4 5 6 7 8 9


Target Classification (L_∞)
0 1 2 3 4 5 6 7 8 9
Source Classification
0 1 2 3 4 5 6 7 8 9

Evaluation metrics – what is a successful adversarial example

- Misclassification
  - *Targeted*: *C(x')=t,* an attacker-chosen target label
  - *Untargeted: C(x') ≠ C(x)*

- Human perceives x' and x as the same
  - Hard to implement – involves user studies
  - Approximate using L_p norm of noise

Other methods

- Beyond L_p norm

- Physically realizable adversarial examples

Beyond images

- Text
- Audio
- Video
- Software

Preserving semantics

C produces an attacker-desired output for *x'*

Formulation as optimization problem