ADVERSARIAL ROBUSTNESS

Recall: ERM: $\underset{\theta}{\text{minimize}}\ \mathbb{E}_{x,y \sim \text{training}}\ \ell(f_\theta(x), y)$

Goal: maximize acc. on test data.

> usually in standard ML, train, test $\overset{iid}{\sim}$

↓ not always true

① DA / DG

ROBUSTNESS → "Adversarial" imperceptible

↓ "Distributional"

$$x + \delta = x'$$

- x: input
- δ: "some" noise
- x': adversarial example

DEFINITION: x' is an adv. version of x (given model $f_\theta(\cdot)$)

if

① $f_\theta(x) \neq f_\theta(x')$ — pred. not same

② $f_{\text{human}}(x) = f_{\text{human}}(x')$ — human pred. identical

$f_{\text{det}}(x) = f_{\text{det}}(x')$ generally

(3) for a human, the diff. bet^n x and x'

$$d_{human}(x, x') = 0$$

↓ assumption

for some distance measure d

$$d(x, x') \leq \epsilon$$

Q: (1) what noise δ should we add?

Q: (2) can we quantify "imperceptible" what is $d(\cdot)$ (What is the THREAT MODEL)

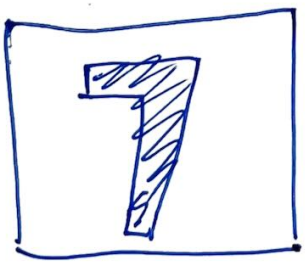
e.g.

L_2 norm	$\|x' - x\|_2 \leq \epsilon$
L_1 norm	$\|x' - x\|_1 \leq \epsilon$
L_∞ norm	$\|x' - x\|_\infty \leq \epsilon$

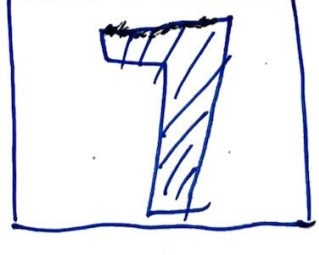
aside: L_∞ norm

$$\|a\|_\infty = \max_i (a_i)$$

L_p norm threat models



rotate it by 0.5°



OPTIMAL TRANSPORT COST

x x'

$$\|x' - x\|_p \leq \epsilon$$

ATTACK : creating adv. examples
(when applied to the model, accuracy drops)

DEFENSE : preventing attacks
improving robustness

ATTACK problem formulation

Goal : given input x, model $f_\theta(\cdot)$
find δ $\quad x' = x + \delta$

s.t. ① $f_\theta(x') \neq f_\theta(x) = y$ → g.t.

② $\|x - x'\|_p \leq \epsilon$

TYPES OF ATTACK BASED ON OUTCOME

① targeted → make the model predict label y_{target}

② untargeted → make the model predict some wrong label

BASED ON WHEN THE MODEL IS ATTACKED

① inference / test time attacks → TODAY

② training time attacks "poisoning"

WHAT DOES THE ATTACKER HAVE ACCESS TO?

① Black Box attack : only access to I/O

② white Box attack : I/O, model weights, architecture ...

ATTACK GOAL : make classifier fail

$$\max_{\delta} \ell(f_\theta(x), y) \quad \text{s.t} \quad \|\delta\|_p \leq \epsilon$$

$$\delta \leftarrow \delta + \eta \nabla \ell$$

- we need to optimize
- we know G.D.
- but constraint ?

METHOD ① Fast Gradient Sign Method (FGSM)

- simple idea
- one-step search
- look @ direction of gradient
- go in that direction
  - ↳ but don't violate the constraint

$$\delta_{FGSM} = \max_{\|\delta\|_\infty < \epsilon} \langle \delta, \nabla_x \ell(f_\theta(x), y) \rangle \quad g_t$$

$$\delta_{FGSM} = \epsilon \cdot \text{sign}\left(\nabla_x \ell(f_\theta(x), y)\right)$$

METHOD (2) Projected G-D

- iterative version FGSM

$$x^0 = x \quad \text{(input)}$$

$$\hat{x}^{t+1} = x^t + \eta\, g^t \longrightarrow \text{from FGSM}$$

problem! what about the constraint

"do a projection"

$$x^{t+1} = \Pi\left(\hat{x}^{t+1}\right)$$

↓ projection

for L_∞ : $x^{t+1} = \text{clip}\left(\hat{x}^{t+1}, x-\epsilon, x+\epsilon\right)$

for L_2 : normalize

ALTERNATIVE FORMULATION

i.e. make attack imperceptible

$$\begin{cases} \text{minimize } \|\delta\|_p \\ \text{s.t. } f_\theta(x+\delta) \neq y \end{cases}$$

eqv.

$$\underset{\delta}{\text{minimize}} \; \|\delta\|_p - \lambda \cdot \ell\left(f_\theta(x+\delta), y\right)$$

ADVERSARIAL DEFENSE

- remember adv. attack is at test-time
- ~~model is already trained~~
- need diff. training method

recall: ERM $\min_\theta \mathbb{E}_{x,y \sim \text{train}} \ell(f_\theta(x), y)$

Adversarial Training

$$\min_\theta \mathbb{E}_{x,y \sim \text{training}} \left[\max_{x'} \left[\ell(f_\theta(x'), y) \right] \right]$$

$x' = x + \delta$ s.t. $\|x - x'\| \leq \epsilon$

- instead of minimizing loss on "clean" training data

(1) find worst perturbations (adv. examples)

(2) minimize the worst-case loss

Inner Max: attack problem (solved using PGD, FGSM, ...)

Outer Min: G-D.

↓ "min-max" problems

Alternating Optimization