Can Adversarial Training Be Manipulated By Non-Robust Features?

Lue Tao[1], Lei Feng[2,3], Hongxin Wei[4]
Jinfeng Yi[5], Sheng-Jun Huang[6], Songcan Chen[6]

[1]Nanjing University, China
[2]Chongqing University, Chongqing, China
[3]RIKEN Center for Advanced Intelligence Project, Japan
[4]Nanyang Technological University, Singapore
[5]JD AI Research, China
[6]Nanjing University of Aeronautics and Astronautics, China

NeurIPS 2022

Adversarial Training

- Adversarial training
  - Improving test robustness by minimizing the adversarial risk

Natural training

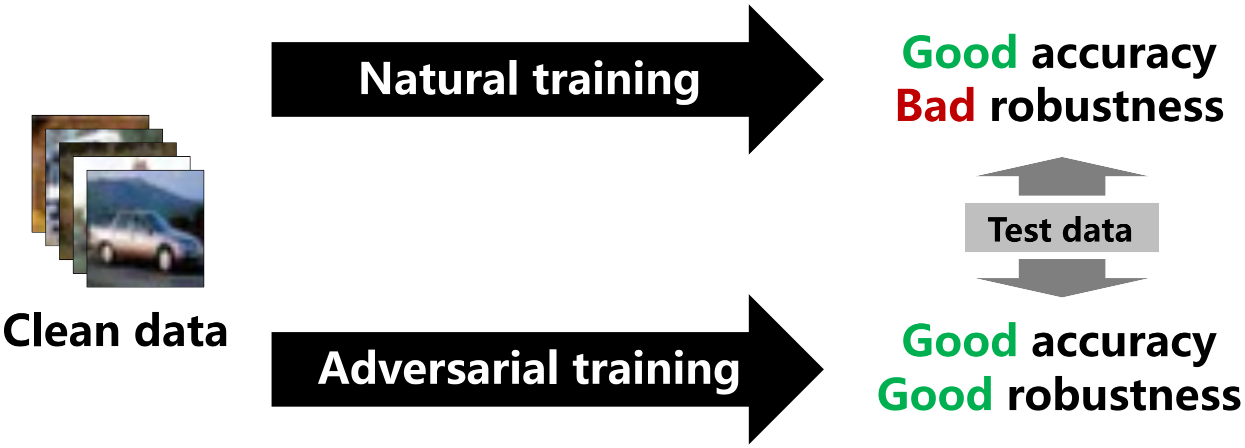
$$\mathcal{R}_{\mathrm{nat}}(f) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathcal{L}(f(\boldsymbol{x}), y)]$$

Adversarial training

$$\mathcal{R}_{\mathrm{adv}}(f) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{\delta}\in\Delta}\mathcal{L}(f(\boldsymbol{x}+\boldsymbol{\delta}), y)\right]$$

Adversarial Training

- Adversarial training
  - Improving test robustness by minimizing the adversarial risk



Adversarial Training

- Adversarial training
  - originally proposed for improving test robustness
  - is capable of mitigating training-time availability attacks

Our Contribution

- We introduce a novel threat model called **stability attack**
  - aims to degrade the test robustness of adversarially trained models
  - in short, aims to hinder robust availability

Our Contribution

- We introduce a novel threat model called **stability attack**
  - aims to degrade the test robustness of adversarially trained models
  - in short, aims to hinder robust availability
- We provide the first theoretical analysis on the robustness of adversarial training against stability attacks
- Comprehensive experiments demonstrate the effectiveness of stability attacks and the necessity of adaptive defense

Theoretical Analysis

- Our binary classification task
  - Gaussian mixture distribution $\mathcal{D}$ $(0 < \eta \ll 1)$

$$y \overset{u.a.r}{\sim} \{-1, +1\}, \quad x_1 \sim \mathcal{N}(y, \sigma^2), \quad x_2, \ldots, x_{d+1} \overset{i.i.d}{\sim} \mathcal{N}(\eta y, \sigma^2)$$

Robust feature **Non-robust features**

- Natural and robust classifiers

$$f_{\text{nat}}(\boldsymbol{x}) := \text{sign}(\boldsymbol{w}_{\text{nat}}^\top \boldsymbol{x}), \text{ where } \boldsymbol{w}_{\text{nat}} := [1, \eta, \ldots, \eta]$$
$$f_{\text{rob}}(\boldsymbol{x}) := \text{sign}(\boldsymbol{w}_{\text{rob}}^\top \boldsymbol{x}), \text{ where } \boldsymbol{w}_{\text{rob}} := [1, 0, \ldots, 0]$$

Theoretical Analysis

- Two representative perturbations
  - Adversarial perturbation
    - shift each feature towards $-y$, resulting in $\mathcal{T}_{\text{adv}}$

$$y \overset{u.a.r}{\sim} \{-1,+1\}, \quad x_1 \sim \mathcal{N}((1-\epsilon)y, \sigma^2), \quad x_2, \ldots, x_{d+1} \overset{i.i.d}{\sim} \mathcal{N}((\eta-\epsilon)y, \sigma^2)$$

  - Hypocritical perturbation
    - shift each feature towards y, resulting in $\mathcal{T}_{\text{hyp}}$

$$y \overset{u.a.r}{\sim} \{-1,+1\}, \quad x_1 \sim \mathcal{N}((1+\epsilon)y, \sigma^2), \quad x_2, \ldots, x_{d+1} \overset{i.i.d}{\sim} \mathcal{N}((\eta+\epsilon)y, \sigma^2)$$

Theoretical Analysis

- Two representative perturbations
  - Adversarial perturbation
    - shift each feature towards $-y$, resulting in $\mathcal{T}_{\text{adv}}$

$$y \overset{u.a.r}{\sim} \{-1, +1\}, \quad x_1 \sim \mathcal{N}((1-\epsilon)y, \sigma^2), \quad x_2, \ldots, x_{d+1} \overset{i.i.d}{\sim} \mathcal{N}((\eta-\epsilon)y, \sigma^2)$$

Theorem 1 (Adversarial perturbation is harmless)**.** *Assume that the adversarial perturbation in the training data $\mathcal{T}_{adv}$ (10) is moderate such that $\eta/2 \leq \epsilon < 1/2$. Then, the optimal linear ℓ_∞-robust classifier obtained by minimizing the adversarial risk on $\mathcal{T}_{adv}$ with a defense budget ϵ is equivalent to the robust classifier (9).*

Theoretical Analysis

- Two representative perturbations
  - Hypocritical perturbation
    - shift each feature towards y , resulting in $\mathcal{T}_{\text{hyp}}$

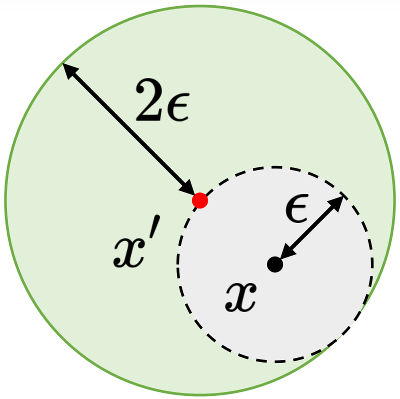
$$y \overset{u.a.r}{\sim} \{-1,+1\}, \quad x_1 \sim \mathcal{N}((1+\epsilon)y, \sigma^2), \quad x_2, \ldots, x_{d+1} \overset{i.i.d}{\sim} \mathcal{N}((\eta+\epsilon)y, \sigma^2)$$

Theorem 2 (Hypocritical perturbation is harmful). *The optimal linear ℓ_∞-robust classifier obtained by minimizing the adversarial risk on the perturbed data $\mathcal{T}_{hyp}$ (11) with a defense budget ϵ is equivalent to the natural classifier (8).*

Theoretical Analysis

- Adaptive defense
  - A defense budget of 2ϵ is capable of resisting any stability attack

Theorem 4 (General case). *For any data distribution and any adversary with an attack budget ϵ, training models to minimize the adversarial risk with a defense budget 2ϵ on the perturbed data is sufficient to ensure ϵ-robustness.*



Theoretical Analysis

- Adaptive defense
  - A defense budget of 2ϵ is capable of resisting any stability attack

Theorem 4 (General case). *For any data distribution and any adversary with an attack budget ϵ, training models to minimize the adversarial risk with a defense budget 2ϵ on the perturbed data is sufficient to ensure ϵ-robustness.*

  - The budget can be reduced to $\epsilon + \eta$ in the Gaussian mixture setting

Theorem 3 ($\epsilon + \eta$ is necessary). *The optimal linear ℓ_∞-robust classifier obtained by minimizing the adversarial risk on the perturbed data $\mathcal{T}_{hyp}$ (11) with a defense budget $\epsilon + \eta$ is equivalent to the robust classifier (9). Moreover, any defense budget lower than $\epsilon + \eta$ will yield classifiers that still rely on all the non-robust features.*

Empirical Evidence

- Stability attacks are harmful to conventional adversarial training

Table 2: Test robustness (%) of PGD-AT using a defense budget $\epsilon_d = 8/255$ on CIFAR-10.

Attack	Natural	FGSM	PGD-20	PGD-100	CW_∞	AutoAttack
None (clean)	82.17	56.63	50.63	50.35	49.37	46.99
DeepConfuse [16]	81.25	54.14	48.25	48.02	47.34	44.79
Unlearnable Examples [28]	83.67	57.51	50.74	50.31	49.81	47.25
NTGA [81]	82.99	55.71	49.17	48.82	47.96	45.36
Adversarial Poisoning [18]	**77.35**	53.93	49.95	49.76	48.35	46.13
Hypocritical Perturbation (ours)	88.07	**47.93**	**37.61**	**36.96**	**38.58**	**35.44**

Empirical Evidence

- Enlarging the defense budget is essential for hypocritical perturbations

Table 5: Test robustness (%) of various adaptive defenses on the hypocritically perturbed CIFAR-10.

Defense	Natural	FGSM	PGD-20	PGD-100	CW_∞	AutoAttack
PGD-AT ($\epsilon_d = 8/255$)	88.07	47.93	37.61	36.96	38.58	35.44
+ Random Noise	87.62	47.46	38.35	37.90	39.07	36.25
+ Gaussian Smoothing	83.95	50.96	42.80	42.34	42.41	40.07
+ Cutout	**88.26**	49.23	39.77	39.25	40.38	37.61
+ AutoAugment	86.24	48.87	40.19	39.65	37.66	35.07
PGD-AT ($\epsilon_d = 14/255$)	80.00	56.86	52.92	52.83	**50.36**	**48.63**
TRADES ($\epsilon_d = 12/255$)	79.63	55.73	51.77	51.63	48.68	47.83
MART ($\epsilon_d = 14/255$)	77.29	**57.10**	**53.82**	**53.71**	49.03	47.67

Summary

- Both theoretical and empirical evidences show that the conventional defense budget ϵ is insufficient under the threat of ϵ-bounded training-time perturbations.
- Our findings suggest that practitioners should consider a larger defense budget of no more than 2ϵ (practically, about $1.5\epsilon \sim 1.75\epsilon$) to achieve a better ϵ-robustness.

Thanks !