Can Adversarial Training Be Manipulated By Non-Robust Features?

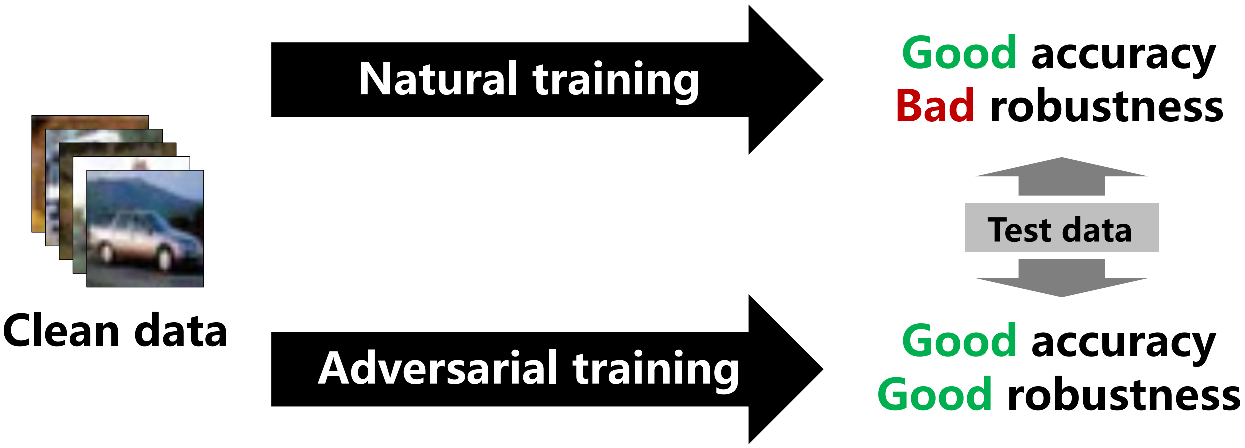
Lue Tao[1], Lei Feng[2,3], Hongxin Wei[4]
Jinfeng Yi[5], Sheng-Jun Huang[6], Songcan Chen[6]

[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



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**,
  - which aims to degrade test robustness of adversarially trained models
  - in short, hinder robust availability

- Both theoretical and empirical evidences show that **adversarial training may fail to provide test robustness**

Theoretical Analysis

Theorem 1 (Adversarial perturbation is harmless). *Assume that the adversarial perturbation in the training data $\mathcal{T}_{adv}$ (10) is moderate such that $\eta/2 \le \epsilon < 1/2$. 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).*

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).*

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.*

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.*

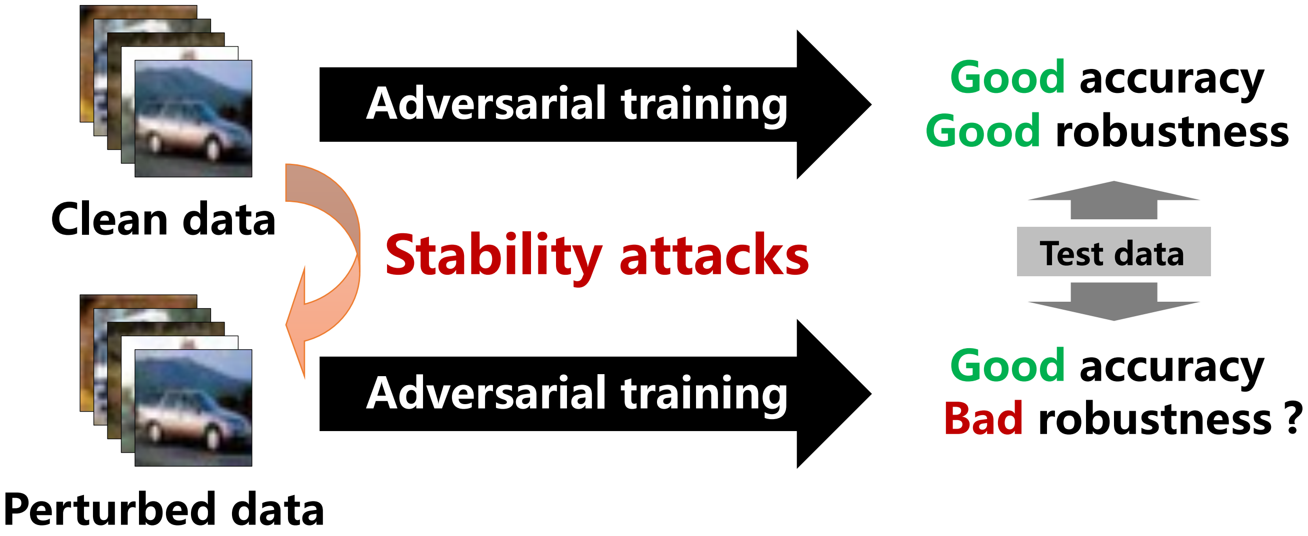
Empirical Evidence

- Stability attacks are harmful to conventional adversarial training
- Enlarging the defense budget is essential for hypocritical perturbations

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**

Summary



Thanks !