Better Safe Than Sorry: Preventing Delusive Adversaries with Adversarial Training

Lue Tao[1,2], Lei Feng[3], Jinfeng Yi[4], Sheng-Jun Huang[1,2], Songcan Chen[1,2*]

[1]Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[3]Chongqing University
[4]JD AI Research

*s.chen@nuaa.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS 2021

Adversarial Examples

Adversarial Examples: worst-case data at **test** time

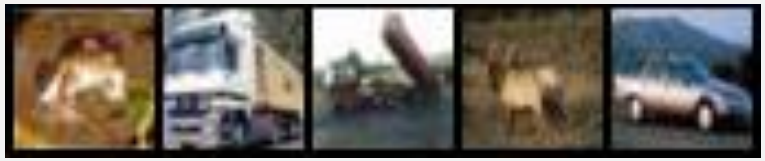
Clean training data $\mathcal{D}$

Clean test data $\mathcal{D}$

Accuracy : 90+%

Perturbations

Worst test data $\mathcal{D}^*$



Accuracy: 0%

[1] Biggio, et al. Evasion attacks against machine learning at test time. ECML-KDD, 2013.
[2] Szegedy, et al. Intriguing properties of neural networks. ICLR, 2014.

What if the training data can be perturbed?

Clean training data $\mathcal{D}$

Perturbations

Worst training data $\widehat{\mathcal{D}}$

Clean test data $\mathcal{D}$

Accuracy : 0%

Delusive Attacks

Delusive Attacks: worst-case data at **training** time

Clean training data $\mathcal{D}$

Perturbations

Worst training data $\widehat{\mathcal{D}}$

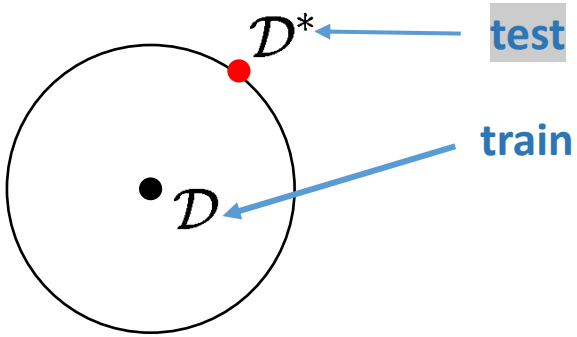
Clean test data $\mathcal{D}$

Accuracy : 0%

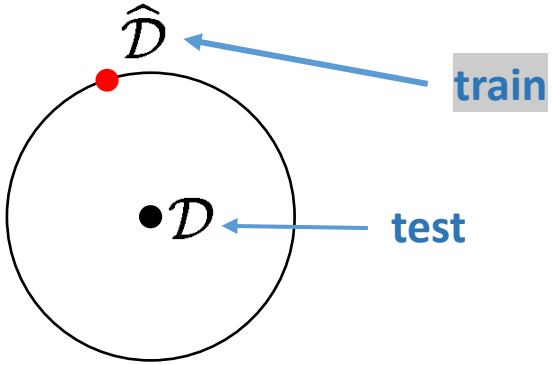
[3] Newsome, et al. Paragraph: Thwarting Signature Learning by Training Maliciously. Recent advances in intrusion detection, 2006.
[4] Feng, et al. Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder. NeurIPS, 2019.

Our Perspective: Twins of Evil

Adversarial Examples: worst-case **test** data

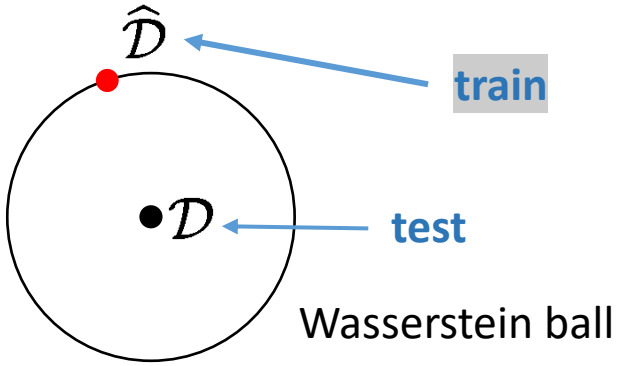


Delusive Attacks: worst-case **training** data



Our Contribution

[Contribution 1] Formulation of delusive attacks



$$\max_{\widehat{\mathcal{D}} \in \mathcal{B}_{\mathrm{W}_\infty}(\mathcal{D}, \epsilon)} \quad \mathop{\mathbb{E}}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\ell(f_{\widehat{\mathcal{D}}}(\boldsymbol{x}), y) \right],$$

$$\text{s.t.} \quad f_{\widehat{\mathcal{D}}} = \arg\min_{f} \mathop{\mathbb{E}}_{(\boldsymbol{x}, y) \sim \widehat{\mathcal{D}}} \left[\ell\left(f\left(\boldsymbol{x}\right), y\right) \right].$$

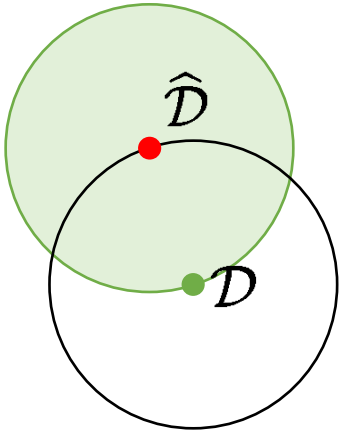
[4] Feng, et al. Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder. NeurIPS, 2019.

Our Contribution

[Contribution 2] The principled defense

Theorem 1. *For any data distribution $\mathcal{D}$ and any delusive distribution $\widehat{\mathcal{D}}$ such that $\widehat{\mathcal{D}} \in \mathcal{B}_{\mathrm{W}_\infty}(\mathcal{D}, \epsilon)$ generated by a delusive adversary, we have*

$$\mathcal{R}_{\mathsf{nat}}(f, \mathcal{D}) \leq \max_{\mathcal{D}' \in \mathcal{B}_{\mathrm{W}_\infty}(\widehat{\mathcal{D}}, \epsilon)} \mathcal{R}_{\mathsf{nat}}(f, \mathcal{D}') = \mathcal{R}_{\mathsf{adv}}(f, \widehat{\mathcal{D}}).$$



Take-aways

1. Minimizing the adversarial risk on the *perturbed data* ⇔ Minimizing an upper bound of natural risk on the *original data*
2. Adversarial Training: A principled defense against delusive attacks

Our Contribution

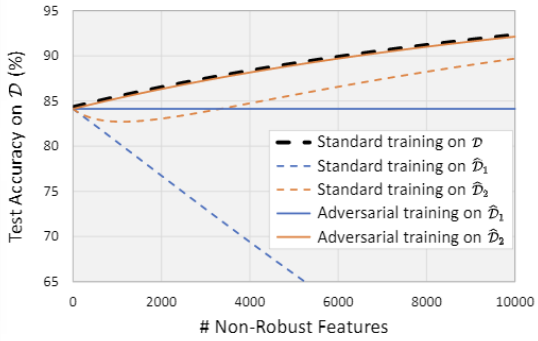
[Contribution 3] Internal Mechanisms

Theorem 2. *Let $f_{\mathcal{D}}$, $f_{\widehat{\mathcal{D}}_1}$, and $f_{\widehat{\mathcal{D}}_2}$ be the Bayes optimal classifiers for the mixture-Gaussian distributions $\mathcal{D}$, $\widehat{\mathcal{D}}_1$, and $\widehat{\mathcal{D}}_2$, defined in Eqs. (5), (6), and (7), respectively. For any $\eta > 0$, we have*

$$\mathcal{R}_{\mathsf{nat}}(f_{\mathcal{D}}, \mathcal{D}) < \mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_2}, \mathcal{D}) < \mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_1}, \mathcal{D}).$$

Theorem 3. *Let $f_{\widehat{\mathcal{D}}_1,\mathsf{rob}}$ and $f_{\widehat{\mathcal{D}}_2,\mathsf{rob}}$ be the optimal linear ℓ_∞ robust classifiers for the delusive distributions $\widehat{\mathcal{D}}_1$ and $\widehat{\mathcal{D}}_2$, defined in Eqs. (6) and (7), respectively. For any $0 < \eta < 1/3$, we have*

$$\mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_1}, \mathcal{D}) > \mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_1,\mathsf{rob}}, \mathcal{D}) \quad \text{and} \quad \mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_2}, \mathcal{D}) > \mathcal{R}_{\mathsf{nat}}(f_{\widehat{\mathcal{D}}_2,\mathsf{rob}}, \mathcal{D}).$$



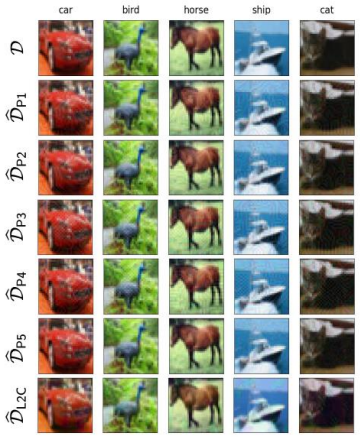
Take-aways

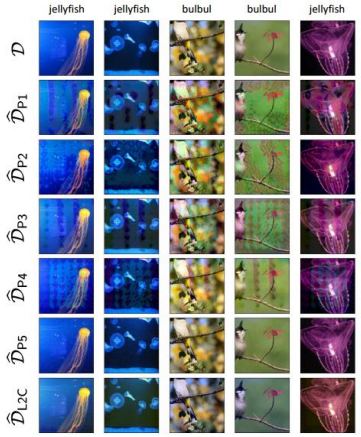
1. Adversarial training works under delusive attacks by mitigating model reliance on non-robust features
2. Adversarial perturbations are more harmful than hypocritical perturbations

Our Contribution

[Contribution 4] Empirical evidences

- Practical delusive attacks
  - Adversarial perturbations (P1),
  - Hypocritical perturbations (P2),
  - Universal adversarial perturbations (P3),
  - Universal hypocritical perturbations (P4),
  - Universal random perturbations (P5),
  - DeepConfuse (L2C) [4]





(a) CIFAR-10 (ℓ_2) (b) SVHN (ℓ_2) (c) CIFAR-10 (ℓ_∞) (d) ImageNet (ℓ_∞) (e) SSL (ℓ_2)

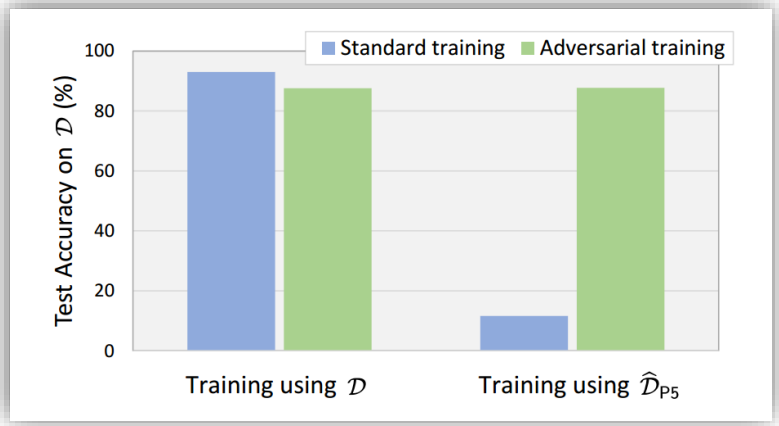
Figure 3: Universal perturbations for the P3 and P4 attacks across different datasets and threat models. Perturbations are rescaled to lie in the $[0, 1]$ range for display. The resulting inputs are nearly indistinguishable from the originals to a human observer (see Appendix B Figures 10, 11, and 12).

[4] Feng, et al. Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder. NeurIPS, 2019.

Our Contribution

[Contribution 4] Empirical evidences

- Six delusive attacks
  Adversarial perturbations (P1), Hypocritical perturbations (P2), Universal adversarial perturbations (P3), Universal hypocritical perturbations (P4), Universal random perturbations (P5), and DeepConfuse (L2C)
- Three datasets
  CIFAR-10, SVHN, and a subset of ImageNet
- Three tasks
  supervised learning, self-supervised learning, and overcoming simplicity bias



Take-aways

1. The defense withstands all the attacks on all the datasets/tasks.
2. Both theoretical and empirical results vote for adversarial training.

Thanks!