Learning to Confuse: Generating Training Time Adversarial Data with Auto-Encoder

Ji Feng*, Qi-Zhi Cai*, Zhi-Hua Zhou

NeurIPS 2019

Presented by Jiaru Zhang, AISIG

March 8, 2022

Shanghai Jiao Tong University

1 Introduction

2 The proposed method

3 Experiments

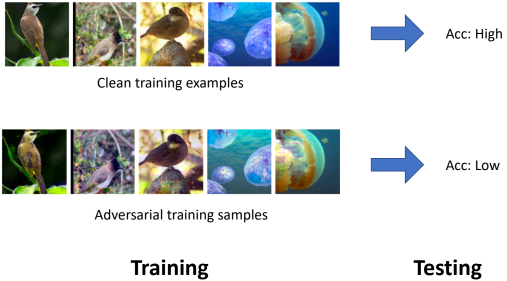
4 Conclusion

Section 1

Introduction

Introduction

- Problem: Adding imperceivable noises to the training data to confuse classifier in testing.



Section 2

The proposed method

Problem formulation

The learning target of a neural network f_θ with parameter θ is

Target

$$\theta^* = \arg\min_\theta \sum_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_\theta(x), y\right)\right] \tag{1}$$

Noise generator: g_ξ

Constraint on noise

$$\forall x, \left\|g_\xi(x)\right\|_\infty \leq \epsilon \tag{2}$$

In this work, an encoder-decoder network with activation $\epsilon \cdot (\tanh(\cdot))$ in the last layer is used.

Problem formulation

The task is formulated into

Task formulation

$$\max_{\xi} \sum_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_{\theta^*(\xi)}(x), y\right)\right]$$
$$\text{s.t.} \quad \theta^*(\xi) = \arg\min_{\theta} \sum_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_{\theta}\left(x + g_{\xi}(x)\right), y\right)\right] \tag{3}$$

SHANGHAI JIAO TONG UNIVERSITY

Optimization

- The equality constraint can be relaxed into

$$\theta_i = \theta_{i-1} - \alpha \cdot \nabla_{\theta_{i-1}} \mathcal{L}\left(f_{\theta_{i-1}}\left(x + g_\xi(x)\right), y\right) \tag{4}$$

- The basic idea is to alternatively update f_θ on **noisy data** via gradient descent.and g_ξ on **clean data** over gradient ascent.
- However, f_θ and g_ξ won't converge in practice.

Optimization

- Collecting the update trajectories for f_θ
- Update g_ξ based on such trajectories.

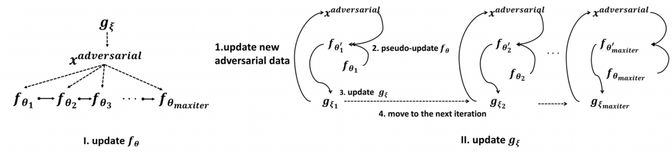


Figure 1: An overview for learning to confuse: Decoupling the alternating update for f_θ and g_ξ

- Implementation trick: save g_ξ instead of f_{θ_i}.

Shanghai Jiao Tong University

Label specific adversaries

- It can be easily transfer to the label specific conditions.

Label specific adversaries

Replace

$$\max_{\xi} \sum_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_{\theta^*}(\xi)(x), y\right)\right] \tag{5}$$

into

$$\min_{\xi} \sum_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_{\theta^*(\xi)}(x), \eta(y)\right)\right], \tag{6}$$

where η is a predefined label transformation function.

Section 3

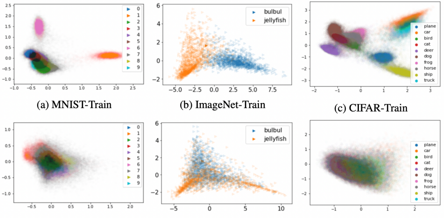
Experiments

Performance Evaluation

- The test accuracy obviously dropped when trained on the adversarial datasets.

	MNIST	ImageNet	CIFAR-10
Clean Data	99.32 ± 0.05	88.5 ± 2.32	77.28 ± 0.17
Adversarial Data	0.25 ± 0.04	54.2 ± 11.19	28.77 ± 2.80

- The classifier trained on the adversarial data cannot differentiate the clean samples.



(a) MNIST-Train (b) ImageNet-Train (c) CIFAR-Train

(d) MNIST-Test (e) ImageNet-Test (f) CIFAR-Test

Shanghai Jiao Tong University

Effect of varying parameters

- There is a sudden drop in performance when the perturbation constraint ϵ exceeds 0.15.
- The proposed method performs better than *random flip*.

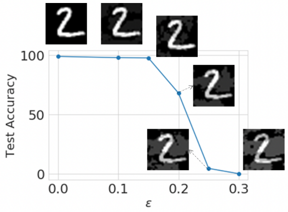


Figure 4: Effect of varying ϵ.

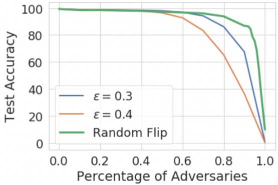
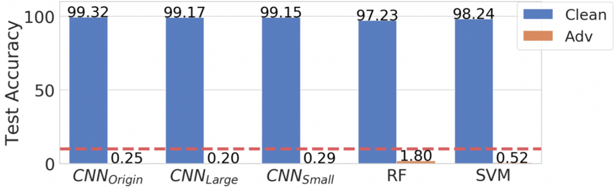


Figure 5: Varying the ratio of adversaries under different ϵ.

Shanghai Jiao Tong University

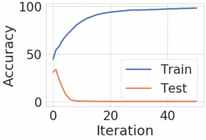
Evaluation of Transferability

- It transfers very well on even non-NN classifiers, e.g., random forest and SVM.

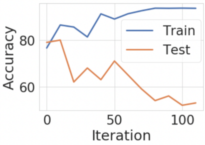


Generalization Gap

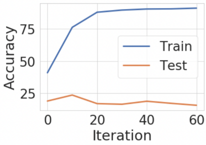
- A clear generalization gap is observed during the training process.
- It is conjectured that the deep model tends to overfit towards the adversarial noises.



(a) MNIST.



(b) 2-class ImageNet.



(c) CIFAR-10.

Figure 8: Learning curves for f_θ

Validation and Linear Hypothesis

- The model performs well when taking only adversarial noises as inputs.
- One possible explanation is the linearity inside deep models.

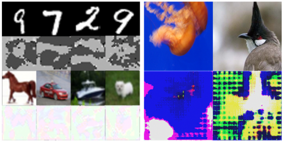


Figure 9: Clean samples and their corresponding adversarial noises for MNIST, CIFAR-10 and ImageNet

Table 2: Prediction accuracy taking **only noises as inputs**. That is, the accuracy between the true label and $f_\theta(g_\xi(x))$ where x is the clean sample.

	$\text{Noise}_{\text{train}}$	$\text{Noise}_{\text{test}}$
MNIST	95.62	95.15
ImageNet	88.87	93.00
CIFAR-10	78.57	72.98

Shanghai Jiao Tong University

Weight Visualizations

- The victim SVM weights went to the opposite direction and tend to overfits on image corners.

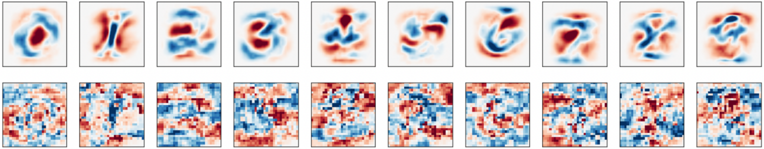
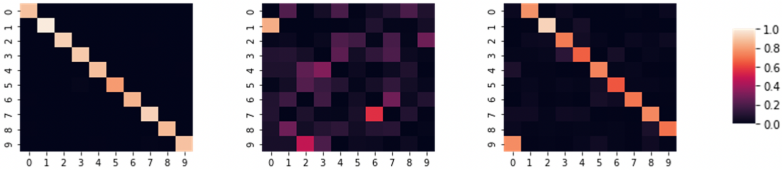


Figure 10: LinearSVM weights visualization for MNIST. Top row: Weights trained on clean training data. Bottom row: Weights trained on adversarial training data.

Label Specific Adversaries

- Price: test accuracy increases from 0.25 ± 0.04 to 1.48 ± 0.21.
- Effect: Success rate for targeting the desired specific label: 79.7 ± 0.38.



(a) Clean Training Data (b) Non-label specific setting (c) Label-specific setting

SHANGHAI JIAO TONG UNIVERSITY

Section 4

Conclusion

Conclusion

- This paper proposed a general framework for generating training time adversarial data.
- A simple yet effective training scheme to train both networks.
- Experiments on image data confirm the effectiveness.

Related consecutive work

- A concurrent work minimizes the gradients of weights to make models harder to converge in transfer learning [1].
- "Inversely adversarial noise" generated by PGD has a similar effect and is used to synthesize *Unlearnable Examples* [2].
- Gradient manipulation is used to generate poisoned dataset [3].
- Adversarial examples make stronger poisons [4].
- Adversarial training serves as a defense with theoretical guarantee [5].

1 Juncheng Shen, Xiaolei Zhu, De Ma. TensorClog: An Imperceptible Poisoning Attack on Deep Neural Network Applications, in IEEE Access, vol. 7, pp. 41498-41506, 2019

2 Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, Yisen Wang. Unlearnable Examples: Making Personal Data Unexploitable. In ICLR, 2021.

3 Liam H Fowl, Ping-yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Amit Bansal, Wojciech Czaja, Tom Goldstein. Protecting Proprietary Data: Poisoning for Secure Dataset Release. In arxiv preprint, 2103.02683.

4 Liam H Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, Tom Goldstein. Adversarial Examples Make Strong Poisons. In NeurIPS, 2021.

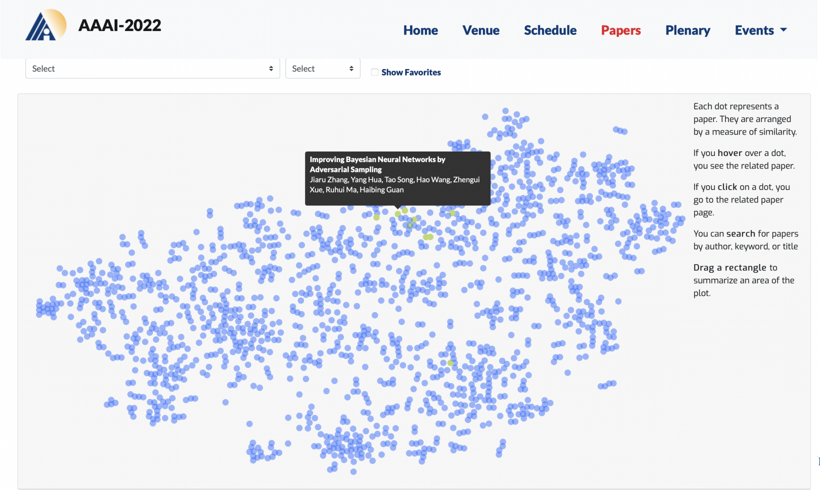
5 Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, Songcan Chen. Better Safe Than Sorry: Preventing Delusive Adversaries with Adversarial Training. In NeurIPS, 2021.

Section 5

Others

Experience in AAAI

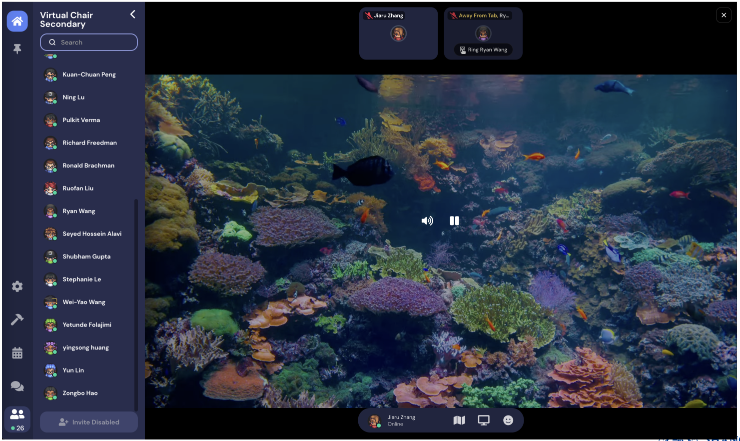
Experience in AAAI



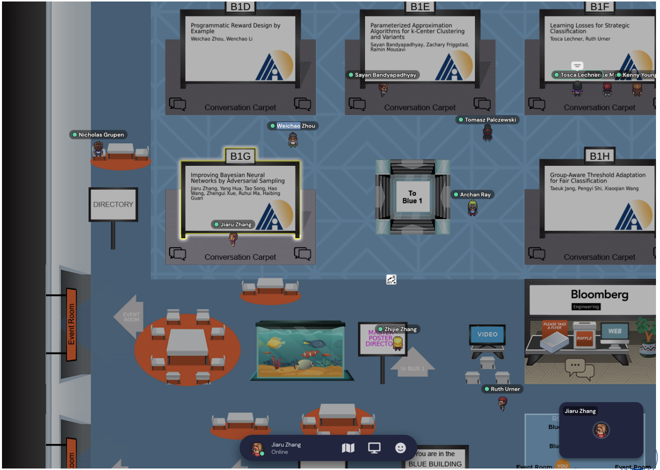
Experience in AAAI

SHANGHAI JIAO TONG UNIVERSITY

Experience in AAAI



Experience in AAAI



SHANGHAI JIAO TONG UNIVERSITY

Experience in AAAI

SHANGHAI JIAO TONG UNIVERSITY