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

Ji Feng[1,2], Qi-Zhi Cai[2] and Zhi-Hua Zhou[1]

1. LAMDA Group, Nanjing University

2. Sinovation Ventures AI Institute

www.chuangxin.com

Motivation and Key Idea

How to perturbate the training data as small as possible so that any trained model performs very badly on clean test?

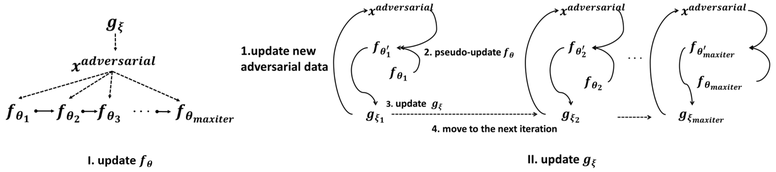
Key Idea: Teaching a perturbation network by hijacking the training process of one imaginary victim differentiable classifier.

Proposed method

We formulate the problem into a non-linear equality constrained optimization problem:

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

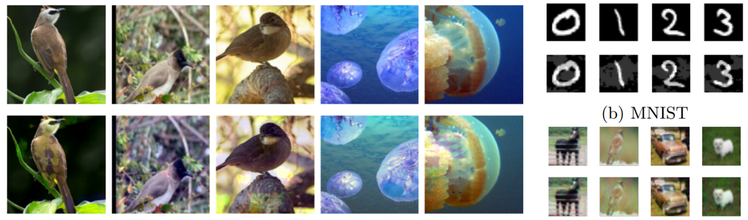
where g_ξ is the noise generator and f_θ is the victim classifier. Every ξ is paired with one classifier $f_{\theta^*(\xi)}$ trained on the corresponding modified data.



Solving such problem is computationally challenging (unlike GANs). We then proposed one simple yet effective procedure to decouple the alternating updates for the two networks for stability. By teaching the perturbation generator to hijacking the training trajectory of the victim classifier, the generator can thus learn to move against the victim classifier step by step. The method proposed in this paper can be easily extended to the label specific setting where the attacker can manipulate the predictions of the victim classifier according to some predefined rules rather than only making wrong predictions.

Proposed Method Cont.

Samples of Adversarial Training Data



(a) 2-Class ImageNet (b) MNIST (c) CIFAR-10

Figure 1: First rows: original training samples. Second rows: adversarial training samples.

Experiments

Performance Evaluation

Table 1: Test accuracy (mean±std) when the classifier is trained on the original clean training set and the adversarial training set,respectively.

	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

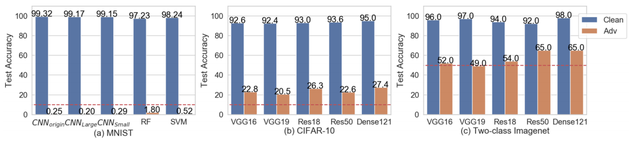


Figure 2: Test performance when using different classifiers. The horizontal red line indicates random guess accuracy.

Experiments Cont.

Transferability

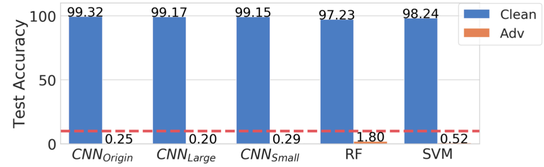
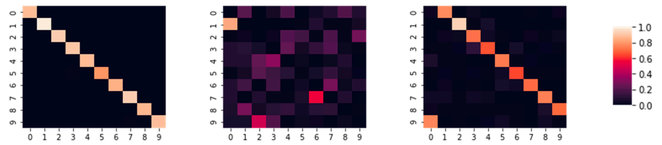


Figure 6: Test performance when using different classifiers. The horizontal red line indicates random guess accuracy.

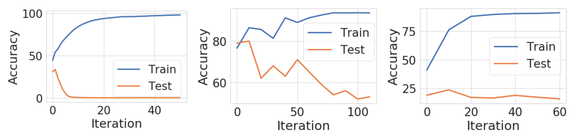
Label Specific Adversaries



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

Figure 11: The confusion matrices on test set under different scenarios for MNIST dataset. They summarized the test performance of classifier trained on (a) clean training data (b) Non-label specific setting and (c) label-specific setting.

The Generalization Gap and Linear Hypothesis



(a) MNIST. (b) 2-class ImageNet. (c) CIFAR-10.

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

Contact: fengji@chuangxin.com