Towards Robust Detection of Adversarial Examples

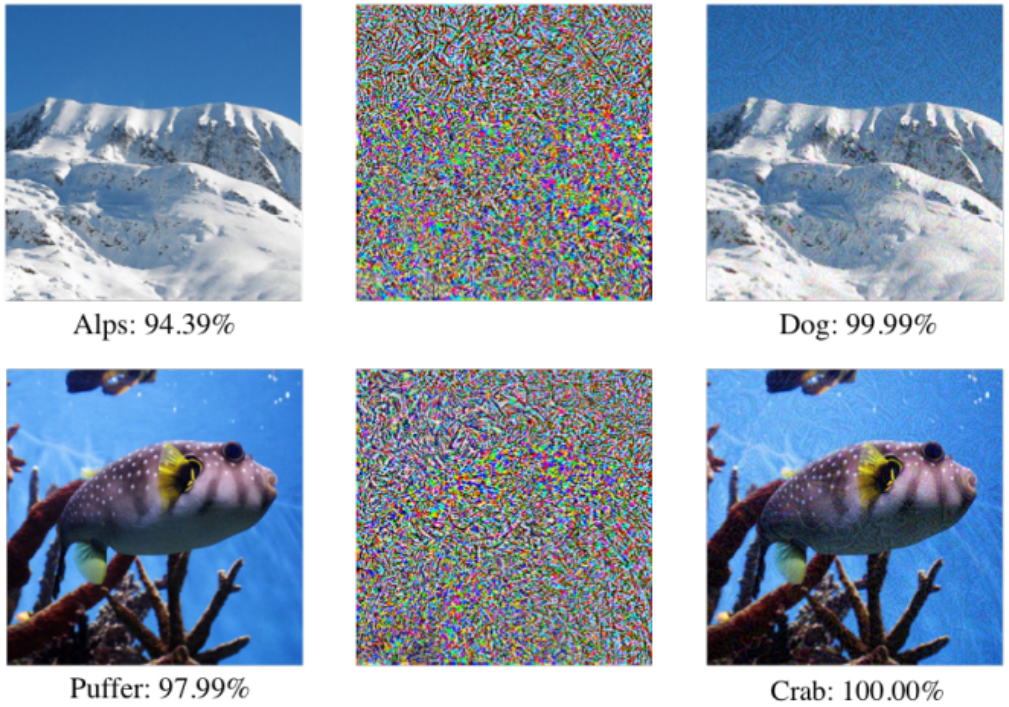
Tianyu Pang, Chao Du, Yinpeng Dong and Jun Zhu

Department of Computer Science and Technology
Tsinghua University

NeurIPS | 2018

TSAIL

Adversarial Examples



From Dong et al. (CVPR 2018)

We Detect Adversarial Examples, and How?

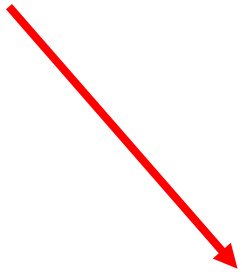
Design new detectors:

- Kernel density detector (Feinman et al. 2017)
- LID detector (Ma et al. ICLR 2018)
- ……

We Detect Adversarial Examples, and How?

Design new detectors:

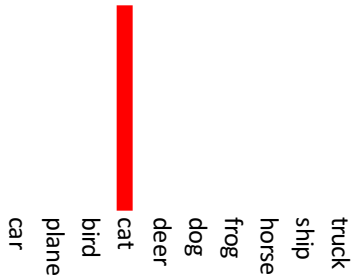
- Kernel density detector (Feinman et al. 2017)
- LID detector (Ma et al. ICLR 2018)
- ……



Train the models to better collaborate with existing detectors

Reverse Cross Entropy

Cross-Entropy (CE):

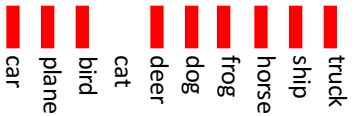


1_y: One-hot label

{0, 0, 0, **1**, 0, 0, 0, 0, 0, 0}

$$\mathcal{L}_{CE} = -\mathbf{1}_y \cdot \log(\mathbf{F})$$

Reverse Cross-Entropy (RCE):



R_y: Reverse label

$\{\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \mathbf{0}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\}$

$$\mathcal{L}_{RCE} = -R_y \cdot \log(\mathbf{F})$$

The RCE Training Method

Phase 1: Reverse Training

Training the model by minimizing the RCE loss

Phase 2: Reverse Logits

Negating the logits fed to the softmax layer to give predictions

Theoretical Analysis

Theorem 2. *(Proof in Appendix A) Let (x, y) be a given training data. Under the L_∞-norm, if there is a training error $\alpha \ll \frac{1}{L}$ that $\|\mathbb{S}(Z_{pre}(x, \theta_R^*)) - R_y\|_\infty \leq \alpha$, then we have bounds*

$$\|\mathbb{S}(-Z_{pre}(x, \theta_R^*)) - 1_y\|_\infty \leq \alpha(L-1)^2,$$

and $\forall j, k \neq y$,

$$|\mathbb{S}(-Z_{pre}(x, \theta_R^*))_j - \mathbb{S}(-Z_{pre}(x, \theta_R^*))_k| \leq 2\alpha^2(L-1)^2.$$

Property 1: Consistent and Unbiased

When the training error $\alpha \longrightarrow 0$, the prediction tends to the one-hot label

Property 2: Tighter Bound

The difference between any two non-maximal elements decreases as $O(\alpha^2)$

Experiments

CE RCE

t-SNE visualization of learned features on CIFAR-10

Experiments

Attack	Obj.	MNIST			CIFAR-10		
		Confidence	non-ME	K-density	Confidence	non-ME	K-density
FGSM	CE	79.7	66.8	98.8 (-)	71.5	66.9	**99.7** (-)
	RCE	98.8	98.6	**99.4** (*)	92.6	91.4	98.0 (*)
BIM	CE	88.9	70.5	90.0 (-)	0.0	64.6	**100.0** (-)
	RCE	91.7	90.6	**91.8** (*)	0.7	70.2	**100.0** (*)
ILCM	CE	98.4	50.4	96.2 (-)	16.4	37.1	84.2 (-)
	RCE	100.0	97.0	**98.6** (*)	64.1	77.8	**93.9** (*)
JSMA	CE	98.6	60.1	97.7 (-)	99.2	27.3	85.8 (-)
	RCE	100.0	99.4	**99.0** (*)	99.5	91.9	**95.4** (*)
C&W	CE	98.6	64.1	99.4 (-)	99.5	50.2	95.3 (-)
	RCE	100.0	99.5	**99.8** (*)	99.6	94.7	**98.2** (*)
C&W-hc	CE	0.0	40.0	91.1 (-)	0.0	28.8	75.4 (-)
	RCE	0.1	93.4	**99.6** (*)	0.2	53.6	**91.8** (*)

AUC-scores (10^{-2}) on adversarial examples

For more results and analyses, please come

Poster: Room 210 & 230 AB #11

Code: https://github.com/P2333/RCE

清華大學
Tsinghua University

NeurIPS | 2018

TSAIL