Towards Robust Detection of Adversarial Examples

Tianyu Pang[1], Chao Du[1], Yinpeng Dong[1] and Jun Zhu[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

NeurIPS | 2018

Backgrounds

Adversarial examples usually have imperceptible perturbation and can fool the model to output wrong predictions. There are typically two strategies to defend adversarial attacks:

- **To correctly classify adversarial examples** (optimal but difficult and sometimes computationally expensive, e.g., adversarial training).
- **To detect adversarial examples** (suboptimal but more computationally efficient, and existing methods can be borrowed from anomaly detection).

Reverse Cross Entropy Training

We detect adversarial examples. Instead of proposing a new detector, we propose a new training method to make networks *better collaborate* with existing detectors like K-density, LID etc. We define the **reverse cross entropy (RCE)**

$$\mathcal{L}^{R}_{CE}(x, y) = -R_y^\top \log F(x), \tag{1}$$

where R_y denote its reverse label vector whose y-th element is zero and other elements equal to $\frac{1}{L-1}$

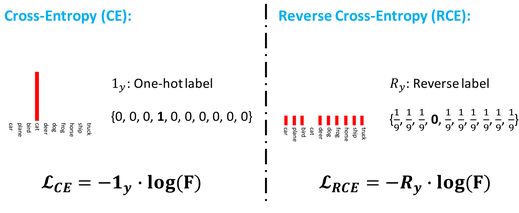

Figure: The difference between the CE and the RCE in the case of $L = 10$.

The RCE training method consists of two simple phases:

- **Reverse training:** Training the DNN $F(X, \theta)$ to be a reverse classifier by minimizing the average RCE loss: $\theta^*_R = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^R_{CE}(x^i, y^i)$.
- **Reverse logits:** Negating the final logits fed to the softmax layer as $F_R(X, \theta^*_R) = \mathbb{S}(-Z_{pre}(X, \theta^*_R))$.

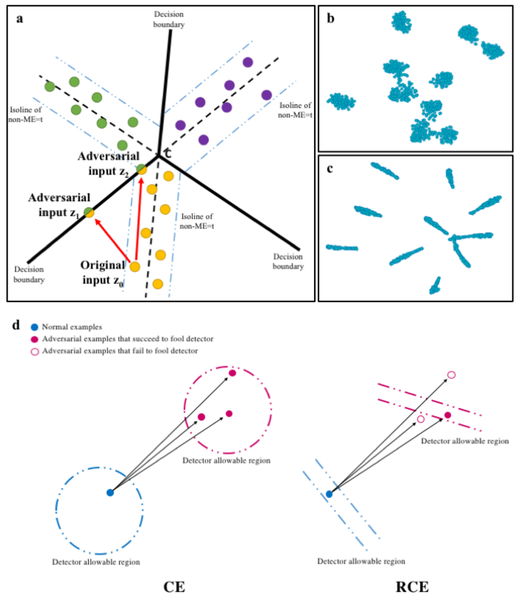

Figure: **a**, The black solid lines are the decision boundary of the classifier. The blue dot-dashed lines are the isolines of non-ME. **b,c**, t-SNE visualization of the learned features (CIFAR-10). **d**, Practical attacks on trained networks.

Theoretical Analyses

We provide some theoretical analyses on RCE training.

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

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

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

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

Theorem 1 demonstrates two important properties of the RCE training procedure.

- **Consistent and unbiased:** When the training error $\alpha \to 0$, the output $F_R(x, \theta^*_R)$ converges to the one-hot label vector 1_y.
- **Tighter bound:** The upper bounds of the difference between any two non-maximal elements in outputs decrease as $\mathcal{O}(\alpha^2)$ w.r.t. α for RCE, much faster than $\mathcal{O}(\alpha)$ for CE and label smoothing.

In the feature space, when non-maximal elements of the prediction tend to be equal, the learned feature will locate near the reverse decision hyperplanes. This will lead to low-dimensional feature distribution for normal examples.

Experiments

We use the kernel density (K-density) applied in the feature space as our detector. The three tables separately show the results under the oblivious, white-box and black-box attacks.

Table: AUC-scores (10^{-2}) of adversarial examples. *Here (-) indicates the baseline, and (*) indicates our method.*

Attack	Obj.	MNIST			CIFAR-10		
		Confi.	non-ME	K-density	Confi.	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** (*)

Obj.	MNIST		CIFAR-10	
	Ratio	Dist.	Ratio	Dist.
CE	0.01	17.12	0.00	1.26
RCE	**0.77**	**31.59**	**0.12**	**3.89**

Table: The ratios of $f_2(x^*) > 0$ and minimal distortions of the adversarial examples crafted by C&W-wb.

	R.-32 (CE)	R.-32 (RCE)
R.-56 (CE)	75.0	90.8
R.-56 (RCE)	89.1	84.9

Table: AUC-scores (10^{-2}) on CIFAR-10. Resnet-32 is the substitute model and Resnet-56 is original.

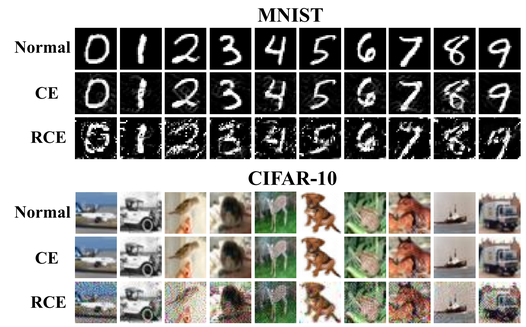

Figure: The normal test images and the adversarial examples generated on Resnet-32 (CE) and Resnet-32 (RCE). Adversarial examples are generated by C&W-wb with minimal distortions. The K-density detector is active.

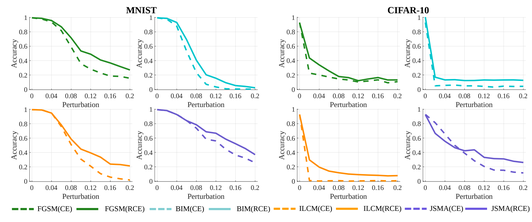

Figure: Classification accuracy under iteration-based attacks on MNIST and CIFAR-10.

Table: Classification error rates (%) on test sets.

Method	MNIST	CIFAR-10
Resnet-32 (CE)	0.38	7.13
Resnet-32 (RCE)	**0.29**	**7.02**
Resnet-56 (CE)	0.36	**6.49**
Resnet-56 (RCE)	**0.32**	6.60

Contact

- **Email**: pty17@mails.tsinghua.edu.cn
- **Code**: https://github.com/P2333/RCE

Thirty-second Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada