Eliminating Catastrophic Overfitting Via Abnormal Adversarial Examples Regularization

Runqi Lin, Chaojian Yu, Tongliang Liu

Sydney AI Centre, The University of Sydney

Single-step Adversarial Training (SSAT)

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \Delta} \ell(x+\delta, y; \theta) \right]$$

Equation 1. The min-max optimization of adversarial training.

Figure 1. The adversarial example generated by SSAT[1].

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Catastrophic Overfitting (CO)

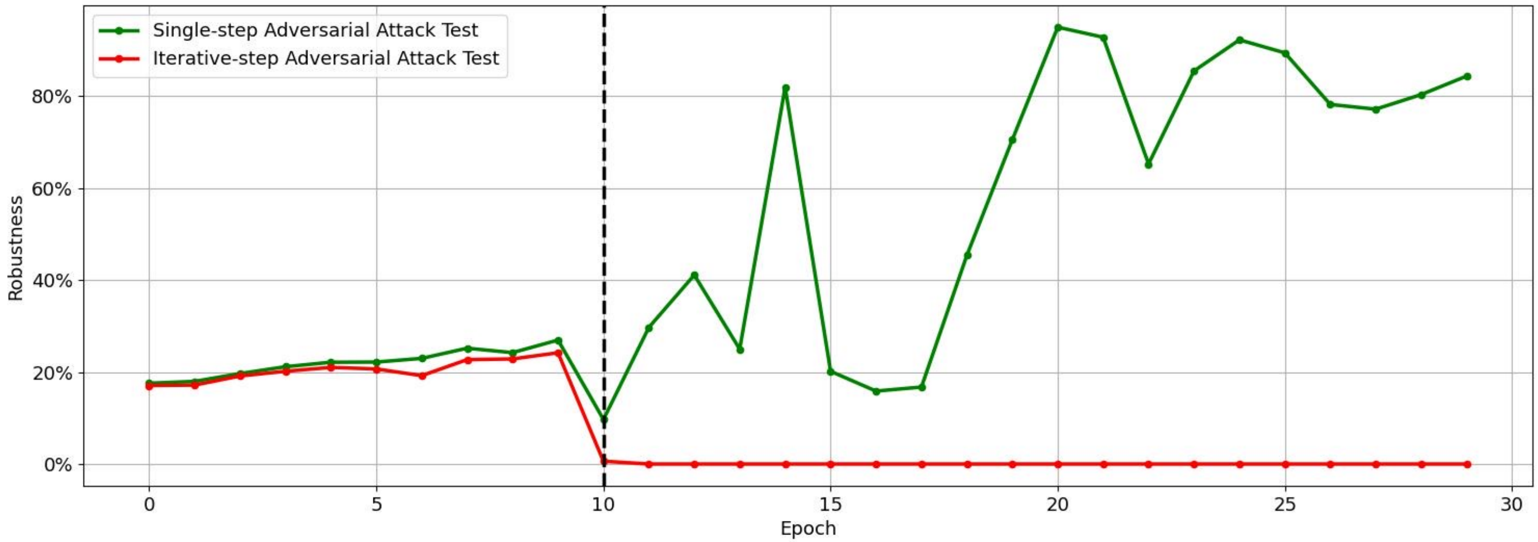

Figure 2. The catastrophic overfitting phenomenon.

Motivation

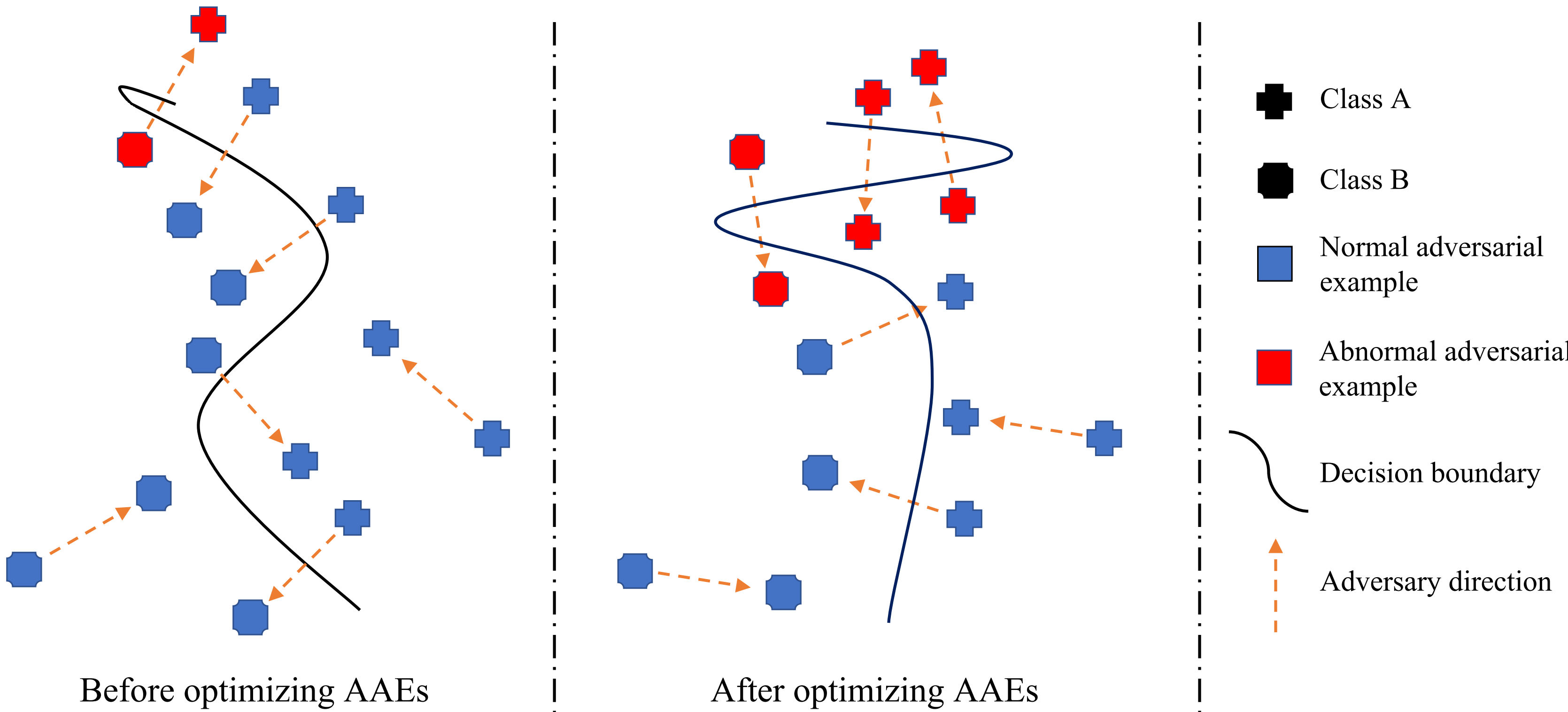

Figure 3. The training samples belonging to NAE (blue) can effectively mislead the classifier, while AAE (red) cannot. The left/middle panel shows the decision boundary before/after optimizing AAEs.

The Definition of Abnormal Adversarial Example (AAE)

$$\delta = \text{sign}\left(\nabla_{x+\eta}\ell(x+\eta, y; \theta)\right),$$

$$x^{AAE} \stackrel{def}{=} \ell\left(x+\eta, y; \theta\right) > \ell\left(x+\eta+\delta, y; \theta\right).$$

Equation 2. The Definition of AAE.

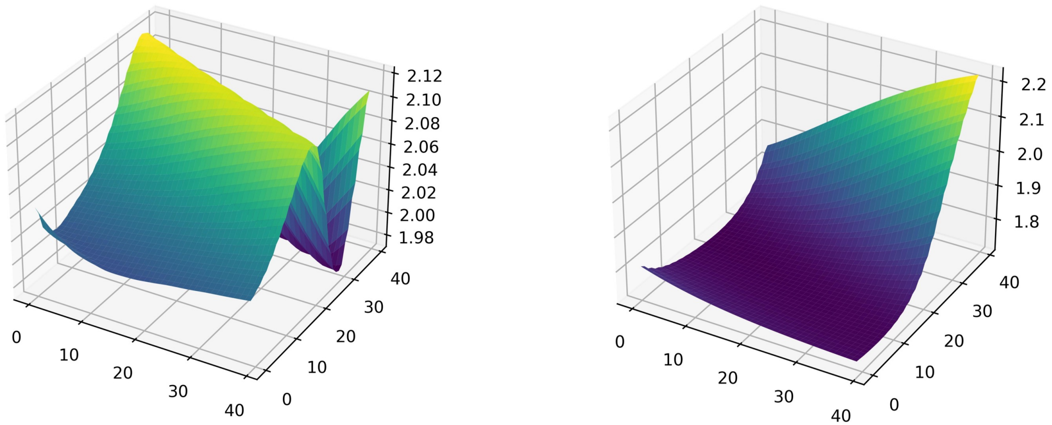

Figure 4. The visualization of AAEs and NAEs loss surface before CO.

Number and Outputs Variation of AAE

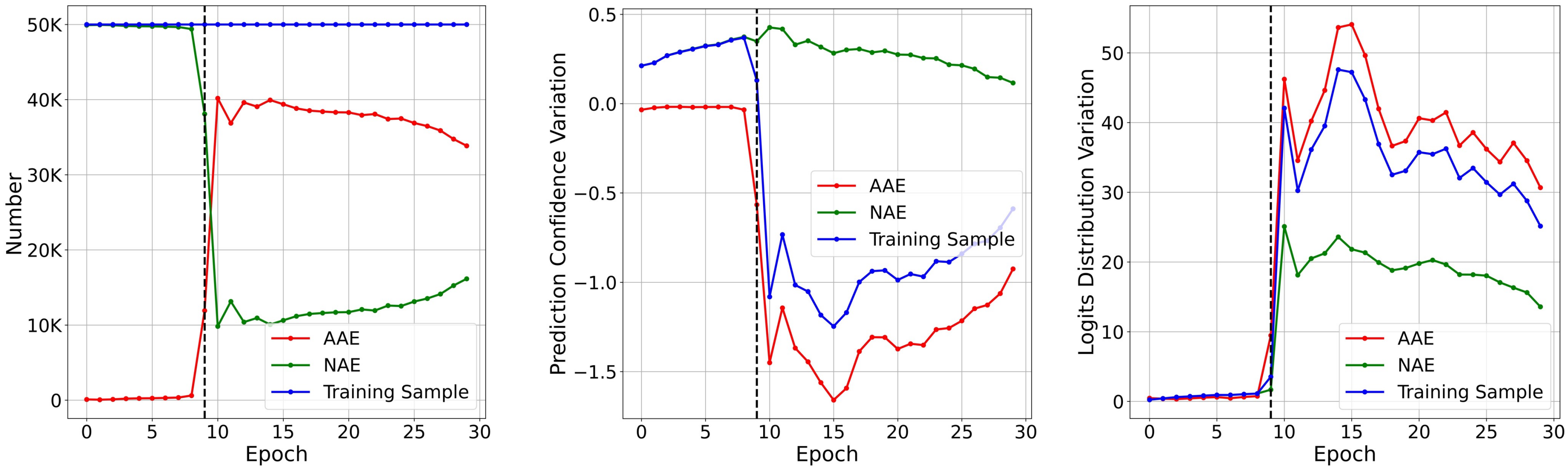

Figure 5. The number, the variation of prediction confidence and logits distribution for NAEs, AAEs and training samples.

Abnormal Adversarial Example Regularization (AAER)

$$AAE\_CE = \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(x_i^{AAE}+\eta, y_i; \theta\right) - \ell\left(x_i^{AAE}+\eta+\delta, y_i; \theta\right)\right).$$

$$AAE\_L2 = \frac{1}{n}\sum_{i=1}^{n}\left(\|f_\theta\left(x_i^{AAE}+\eta+\delta\right) - f_\theta\left(x_i^{AAE}+\eta\right)\|_2^2\right);$$

$$NAE\_L2 = \frac{1}{m-n}\sum_{j=1}^{m-n}\left(\|f_\theta\left(x_j^{NAE}+\eta+\delta\right) - f_\theta\left(x_j^{NAE}+\eta\right)\|_2^2\right).$$

$$AAER = \left(\frac{n}{m}\cdot\lambda_1\right)\cdot\left(AAE\_CE\cdot\lambda_2 + max\left(AAE\_L2 - NAE\_L2, 0\right)\cdot\lambda_3\right).$$

Equation 3. The optimization objectives of AAER: the number, prediction confidence and logits distribution of AAEs.

Experiments

Table 1. Comparison with competing baselines on CIFAR-10/100 datasets.

dataset	CIFAR10				CIFAR100			
noise magnitude	8/255	12/255	16/255	32/255	8/255	12/255	16/255	32/255
FreeAT	76.20 ± 1.09	68.07 ± 0.38	45.84 ± 19.07	61.11 ± 8.41	47.41 ± 0.30	39.84 ± 0.40	3.32 ± 2.48	26.2 ± 15.54
	43.74 ± 0.41	33.14 ± 0.62	0.00 ± 0.00	0.00 ± 0.00	22.27 ± 0.33	16.57 ± 0.20	0.00 ± 0.00	0.00 ± 0.00
ZeroGrad	81.60 ± 0.16	77.52 ± 0.21	79.65 ± 0.17	65.48 ± 6.26	53.83 ± 0.22	49.07 ± 0.14	50.76 ± 0.02	49.38 ± 1.39
	47.56 ± 0.16	27.34 ± 0.09	6.37 ± 0.23	0.00 ± 0.00	25.02 ± 0.24	14.76 ± 0.26	5.23 ± 0.09	0.00 ± 0.00
MultiGrad	81.65 ± 0.16	81.09 ± 4.67	82.98 ± 3.30	70.84 ± 4.53	53.11 ± 0.34	46.81 ± 0.51	46.05 ± 8.68	28.33 ± 6.48
	47.93 ± 0.18	9.95 ± 16.97	0.00 ± 0.00	0.00 ± 0.00	25.68 ± 0.21	16.56 ± 0.56	0.00 ± 0.00	0.00 ± 0.00
Grad Align	82.10 ± 0.78	74.17 ± 0.55	60.37 ± 0.95	25.23 ± 3.41	54.00 ± 0.44	45.83 ± 0.72	36.80 ± 0.10	15.05 ± 0.07
	47.77 ± 0.58	34.87 ± 1.00	27.90 ± 1.01	11.53 ± 3.23	25.27 ± 0.68	18.13 ± 0.71	13.77 ± 0.76	2.85 ± 1.34
RS-FGSM	83.91 ± 0.21	66.46 ± 22.80	66.54 ± 12.25	36.43 ± 7.86	60.29 ± 1.51	18.19 ± 8.51	11.03 ± 5.24	11.40 ± 8.60
	46.01 ± 0.18	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	10.58 ± 13.10	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
N-FGSM	80.48 ± 0.21	71.30 ± 0.12	62.96 ± 0.74	29.79 ± 3.87	54.92 ± 0.28	46.16 ± 0.13	37.93 ± 0.22	18.18 ± 4.55
	47.91 ± 0.29	36.23 ± 0.10	27.14 ± 1.44	8.30 ± 7.85	26.29 ± 0.41	18.75 ± 0.19	14.05 ± 0.07	0.00 ± 0.00
RS-AAER	83.83 ± 0.27	74.40 ± 0.79	64.56 ± 1.45	31.58 ± 1.13	57.71 ± 0.29	44.06 ± 0.93	33.10 ± 0.05	18.50 ± 1.68
	46.14 ± 0.02	32.17 ± 0.16	23.87 ± 0.36	10.62 ± 0.51	25.31 ± 0.01	16.41 ± 0.13	11.80 ± 0.17	4.90 ± 0.50
N-AAER	80.56 ± 0.35	71.15 ± 0.18	61.84 ± 0.43	27.08 ± 0.02	54.47 ± 0.45	45.98 ± 0.13	36.80 ± 0.14	16.95 ± 0.44
	48.31 ± 0.23	**36.52 ± 0.10**	**28.20 ± 0.71**	**12.97 ± 0.57**	**26.81 ± 0.13**	**19.03 ± 0.04**	**14.31 ± 0.05**	**5.45 ± 0.14**
PGD-2	85.07 ± 0.12	78.97 ± 0.23	72.31 ± 0.40	48.45 ± 0.71	60.09 ± 0.20	53.46 ± 0.27	47.50 ± 0.28	31.89 ± 0.69
	45.27 ± 0.07	32.99 ± 0.46	24.32 ± 0.64	11.24 ± 0.40	24.58 ± 0.12	17.16 ± 0.21	12.69 ± 0.06	4.51 ± 0.21
PGD-10 (20)	80.55 ± 0.37	72.37 ± 0.31	67.20 ± 0.69	34.70 ± 0.67	55.05 ± 0.25	47.42 ± 0.29	42.39 ± 0.17	21.68 ± 0.18
	50.67 ± 0.40	**38.60 ± 0.39**	**29.34 ± 0.18**	**16.10 ± 0.20**	**27.87 ± 0.12**	**20.29 ± 0.18**	**15.01 ± 0.21**	**7.39 ± 0.38**

Experiments

Table 2. Comparison with competing baselines on computational overhead.

Method	FreeAT	ZeroGrad	MultiGrad	Grad Align	RS/N-FGSM	RS/N-AAER	PGD-2	PGD-10
Training Time (S)	43.8	11.0	21.7	36.1	11.0	11.2	16.4	59.1

Table 3. Comparison with competing baselines on WideResNet-34 architecture.

method	RS-FGSM	N-FGSM	RS-AAER	N-AAER	PGD-2	PGD-10
natural accuracy (%)	84.41 ± 0.45	84.67 ± 0.32	87.39 ± 0.14	84.47 ± 0.23	88.68 ± 0.14	85.53 ± 0.22
robust accuracy (%)	0.00 ± 0.00	49.72 ± 0.25	47.58 ± 0.42	**50.07 ± 0.53**	47.32 ± 0.50	**53.70 ± 0.53**
training time (S)	98.2		98.6		147.1	536.2

Experiments

Table 4. Comparison with competing baselines on the ImageNet-100 dataset.

method	RS-FGSM	N-FGSM	RS-AAER	N-AAER
natural accuracy (%)	27.10 ± 11.44	38.87 ± 0.17	32.28 ± 1.52	39.52 ± 0.42
robust accuracy (%)	0.00 ± 0.00	20.71 ± 0.74	14.22 ± 0.96	**20.90 ± 0.34**

Table 5. Comparison with competing baselines on the long training schedule.

method	RS-FGSM	N-FGSM	RS-AAER	N-AAER
natural accuracy (%)	91.21 ± 0.26	83.25 ± 0.04	85.69 ± 0.20	83.23 ± 0.25
robust accuracy (%)	0.13 ± 0.02	36.98 ± 0.34	36.05 ± 0.17	**37.38 ± 0.16**

Project Page