Boosting Adversarial Training with Hypersphere Embedding

Tianyu Pang*, Xiao Yang*, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su

Department of Computer Science and Technology, Tsinghua University, Beijing, China

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Adversarial examples cast potential risks when applying machine learning models. Among existing defenses, adversarial training methods can achieve sate-of-the-art performance under different tasks and settings.

Methodology

General framework of adversarial training (AT):

$$\min_{\boldsymbol{\omega},\mathbf{W}} \mathbb{E}\left[\mathcal{L}_{\mathrm{T}}(\boldsymbol{\omega},\mathbf{W}|x,x^*,y)\right],$$
$$\text{where } x^* = \arg\max_{x'\in\mathbf{B}(x)} \mathcal{L}_{\mathrm{A}}(x'|x,y,\boldsymbol{\omega},\mathbf{W}).$$

Affine mapping in the softmax layer:

$$\mathbf{W}^\top z = (W_1^\top z, \cdots, W_L^\top z)$$

Hypersphere embedding (HE):

WN operation: $\widetilde{W}_l = \dfrac{W_l}{\|W_l\|}$

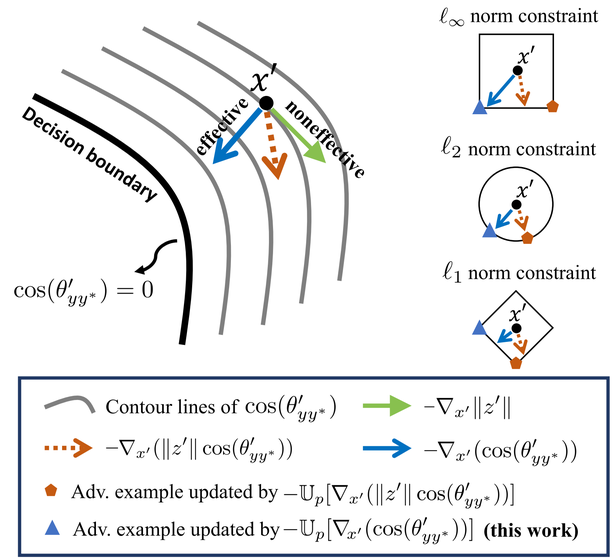
FN operation: $\widetilde{z} = \dfrac{z}{\|z\|}$

AM operation

$$\mathcal{L}_{\mathrm{CE}}^m(\widetilde{f}(x),y) = -1_y^\top \log \mathbb{S}(s\cdot(\cos\boldsymbol{\theta} - m\cdot 1_y))$$

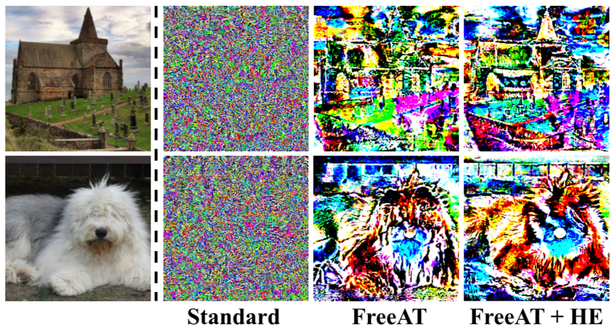
Why HE Benefits AT?

- More effective adversarial perturbations (FN)
- Better learning on hard adversarial examples (FN)
- Alleviate class imbalance caused by untargeted AT (WN)
- Increase inter-class margin (AM)

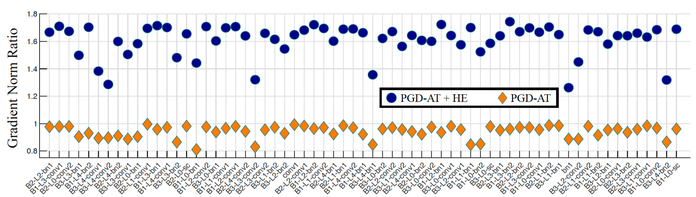


Empirical Results

Visualization of adversarial perturbations on ImageNet:



Gradient norm ratios on adv and clean inputs:



Classification accuracy on CIFAR-10 (C) and ImageNet (C):

Table 2: Classification accuracy (%) on **CIFAR-10** under the *white-box* threat model. The perturbation $\epsilon = 0.031$, step size $\eta = 0.003$. We highlight the best-performance model under each attack.

Defense	Clean	PGD-20	PGD-500	MIM-20	FGSM	DeepFool	C&W	FeaAtt.	FAB
PGD-AT	86.75	53.97	51.63	55.08	59.70	57.26	84.00	52.38	51.23
PGD-AT+**HE**	86.19	59.36	57.59	60.19	63.77	61.56	84.07	52.88	**54.45**
ALP	87.18	52.29	50.13	53.35	58.99	59.40	84.96	49.55	50.54
ALP+**HE**	**89.91**	57.69	51.78	58.63	65.08	**65.19**	**87.86**	48.64	51.86
TRADES	84.62	56.48	54.84	57.14	61.02	60.70	81.13	55.09	53.58
TRADES+**HE**	84.88	**62.02**	**60.75**	**62.71**	**65.69**	60.48	81.44	**58.13**	53.50

Table 3: Validation of combining FastAT and FreeAT with HE and m-HE on **CIFAR-10**. We report the accuracy (%) on clean and PGD, as well as the total training time (min).

Defense	Epo.	Clean	PGD-50	Time
FastAT	30	**83.80**	46.40	*11.38*
FastAT+**HE**	30	82.58	52.55	*11.48*
FastAT+**m-HE**	30	83.14	**53.49**	*11.49*
FreeAT	10	77.21	46.14	*15.78*
FreeAT+**HE**	10	76.85	50.98	*15.87*
FreeAT+**m-HE**	10	**77.59**	**51.85**	*15.91*

Table 4: Top-1 classification accuracy (%) on **ImageNet** under the *white-box* threat model.

Model	Method	Clean	PGD-10	PGD-50
ResNet-50	FreeAT	60.28	32.13	31.39
	FreeAT+**HE**	**61.83**	**40.22**	**39.85**
ResNet-152	FreeAT	65.20	36.97	35.87
	FreeAT+**HE**	**65.41**	**43.24**	**42.60**
WRN-50-2	FreeAT	64.18	36.24	35.38
	FreeAT+**HE**	**65.28**	**43.83**	**43.47**
WRN-101-2	FreeAT	66.15	39.35	38.23
	FreeAT+**HE**	**66.37**	**45.35**	**45.04**

Table 5: Top-1 classification accuracy (%) on **CIFAR-10-C** and **ImageNet-C**. The models are trained on the original datasets CIFAR-10 and ImageNet, respectively. Here 'mCA' refers to the mean accuracy averaged on different corruptions and severity. Full version of the table is in Appendix C.6.

Defense	mCA	*Blur*				*Weather*				*Digital*			
		Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contra	Elastic	Pixel	JPEG
CIFAR-10-C													
PGD-AT	77.23	81.84	79.69	77.62	80.88	81.32	77.95	61.70	84.05	44.55	80.79	84.76	84.35
PGD-AT+**HE**	**77.29**	81.86	79.45	78.17	80.87	80.77	77.98	62.45	83.67	45.11	80.69	84.16	84.10
ALP	77.73	81.94	80.31	78.23	80.97	81.74	79.26	61.51	84.88	45.86	80.91	85.09	84.68
ALP+**HE**	**80.55**	80.87	85.23	81.26	84.43	85.14	83.89	68.83	88.33	50.74	84.44	87.44	87.28
TRADES	75.36	79.84	77.72	76.34	78.66	79.52	76.94	59.68	82.06	43.80	78.53	82.65	82.31
TRADES+**HE**	**75.78**	80.55	77.61	77.26	79.62	79.23	76.53	61.39	82.33	45.04	79.29	82.50	82.40
ImageNet-C													
FreeAT	28.22	19.15	26.63	25.75	28.25	23.03	23.47	3.71	45.18	5.40	41.76	48.78	52.55
FreeAT+**HE**	**30.04**	21.16	29.28	28.08	30.76	26.62	28.35	5.34	49.88	7.03	44.72	51.17	55.05

Table 6: Classification accuracy (%) on the clean test data, and under two benchmark attacks RayS and AutoAttack.

Method	Architecture	Clean	RayS	AA
PGD-AT+**HE**	WRN-34-10	86.25	57.8	53.16
	WRN-34-20	85.14	59.0	53.74

Table 7: Attacking standardly trained WRN-34-10 with or without FN.

Attack	FN	Acc. (%)
PGD-1	✖	67.09
	✔	**62.89**
PGD-2	✖	50.37
	✔	**33.75**

Primary Contacts: Tianyu Pang (pty17@mails.tsinghua.edu.cn), Xiao Yang (yangxiao19@mails.tsinghua.edu.cn); Code: github.com/ShawnXYang/AT_HE