Focus on Hiders: Exploring Hidden Threats for Enhancing Adversarial Training

Qian Li[1,2], Yuxiao Hu[2,3], Yinpeng Dong[4], Dongxiao Zhang[2], Yuntian Chen[2*]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3]The Hong Kong Polytechnic University, HongKong, China
[4]Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, THBI Lab, BNRist Center, Tsinghua University, Beijing 100084, China
qianl01205@sjtu.edu.cn, huyuxiao20@mails.ucas.ac.cn
dongyinpeng@mail.tsinghua.edu.cn, {dzhang, ychen}@eitech.edu.cn

1. Proof of Theorem 1

Proof. Let $\hat{\boldsymbol{x}}^* = \boldsymbol{x} + \hat{\boldsymbol{\delta}}^*$ denotes the worst-case hider of sample $(\boldsymbol{x}, \boldsymbol{y})$ at the i-th epoch, $\hat{\boldsymbol{\delta}}^* \in \mathcal{B}(\epsilon) \cap S_i$ and $\hat{\boldsymbol{x}}^*$ has the highest loss value at the j-th epoch. S_i denotes the perturbation set that every $\hat{\boldsymbol{\delta}} \in S_i$ make $\hat{\boldsymbol{x}} = \boldsymbol{x} + \hat{\boldsymbol{\delta}}$ within the decision boundary at the i-th epoch.

(1) If $j = i + 1$, then the theorem follows naturally.

(2) If $j > i + 1$, then there must exist an epoch l, $i + 1 \leq l < j$, satisfy that $\hat{\boldsymbol{\delta}}^* \in S_l$ at the l-th epoch, and $\hat{\boldsymbol{\delta}}^* \notin S_{l+1}$ at the $(l + 1)$-th epoch, *i.e.*, $\hat{\boldsymbol{x}}^*$ is not an adversarial example at the l-th epoch, but becomes an adversarial example at the $(l + 1)$-th epoch. If $\hat{\boldsymbol{x}}^*$ is the worst-case hider at the l-th epoch, then we can optimize the $\hat{\boldsymbol{x}}^*$ at the l-th epoch. If $\hat{\boldsymbol{x}}^*$ is not the worst-case hider at the l-th epoch, which suggests that $\hat{\boldsymbol{x}}^*$ no longer has the highest upper bound on its attack performance during the future epochs, then we can indicate $\hat{\boldsymbol{x}}^*$ has been indirectly defended between the i-th and l-th epochs.

□

2. Performance on robustness and accuracy of CIFAR-100 and SVHN

Table 1. Comparison of performance improvement of HFAT applied to five different baselines on the CIFAR-100 and implemented them on the PreAct ResNet-18 and WideResNet34-10 architectures.

	PreAct ResNet-18								WideResNet34-10							
	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA
AT_{PGD}	55.57	31.57	28.79	28.66	26.77	28.91	25.71	24.44	58.81	33.75	30.78	30.63	29.41	30.80	27.11	25.84
AT_{HF}	**57.45**	**34.91**	**32.32**	**32.31**	**29.43**	**32.38**	**28.51**	**27.15**	**58.99**	**37.07**	**34.47**	**34.43**	**31.18**	**34.57**	**30.24**	**28.65**
TRADES	54.61	32.86	30.10	29.97	27.24	30.18	26.18	25.81	55.54	34.77	32.86	32.81	31.37	31.88	28.57	27.36
$TRADES_{HF}$	**55.75**	**34.96**	**32.29**	**32.19**	**29.06**	**31.41**	**27.67**	**27.00**	**58.70**	**35.59**	**34.49**	**34.45**	**32.63**	**32.49**	**31.61**	**30.29**
MART	53.41	32.85	30.53	30.39	27.93	29.77	26.12	25.31	53.84	34.42	31.62	31.51	30.14	31.64	28.27	27.08
$MART_{HF}$	**54.74**	**35.02**	**32.78**	**32.55**	**29.82**	**30.62**	**27.93**	**27.51**	**56.19**	**35.31**	**33.52**	**33.39**	**31.40**	**32.22**	**31.49**	**30.34**
AWP	54.19	33.21	30.71	30.62	28.05	29.93	26.54	25.49	54.58	34.03	32.24	32.04	30.67	31.53	29.63	28.75
AWP_{HF}	**55.37**	**35.21**	**33.14**	**33.01**	**30.16**	**31.56**	**27.61**	**27.85**	**57.16**	**35.11**	**33.02**	**32.94**	**31.09**	**33.06**	**31.87**	**31.20**
HELP	54.17	33.56	31.15	30.96	28.42	29.86	26.17	25.61	55.32	34.47	31.54	31.51	29.75	31.65	29.33	28.19
$HELP_{HF}$	**55.23**	**35.46**	**33.65**	**33.32**	**30.24**	**32.07**	**28.26**	**27.87**	**58.05**	**36.40**	**34.42**	**34.38**	**32.58**	**33.27**	**32.58**	**31.36**

Based on the Tab. 1 and Tab. 2, as well as the relevant result on the CIFAR-10 dataset in the main text, it can be demonstrated that HFAT achieves significant improvements across different datasets, defense methods, and network architectures, thus verifying the generality of the strategy.

3. Learning curves

In Fig. 1, we show the learning curve of HFAT, and we find that focusing on hidden threats can effectively prevent overfitting in adversarial training.

*Corresponding author

Table 2. Comparison of performance improvement of HFAT applied to five different baselines on the SVHN and implemented them on the PreAct ResNet-18 and WideResNet34-10 architectures.

	PreAct ResNet-18								WideResNet34-10							
	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA	Natural	FGSM	PGD_{20}	PGD_{100}	CW	MIM	AA_{rand}	AA
AT_{PGD}	90.46	61.62	52.79	50.19	35.03	37.64	34.60	34.10	92.06	64.75	55.26	53.17	36.32	38.87	36.46	35.81
AT_{HF}	**92.56**	**67.69**	**59.12**	**57.03**	**42.47**	**45.50**	**41.64**	**40.83**	**93.16**	**68.21**	**60.46**	**57.92**	**43.41**	**46.79**	**43.70**	**42.88**
TRADES	88.23	63.90	58.26	56.91	36.12	41.56	36.02	35.52	88.90	66.49	59.27	57.23	37.72	43.18	38.21	37.31
$TRADES_{HF}$	**91.31**	**68.74**	**61.13**	**59.72**	**41.37**	**46.94**	**41.05**	**40.56**	**90.73**	**69.30**	**62.23**	**60.33**	**43.65**	**47.34**	**43.35**	**41.18**
MART	89.13	63.19	57.46	55.40	35.84	41.67	35.62	35.15	89.67	66.36	58.64	57.51	36.44	43.52	37.96	37.12
$MART_{HF}$	**90.71**	**67.64**	**60.21**	**58.37**	**40.18**	**47.07**	**40.93**	**40.10**	**91.02**	**69.45**	**62.41**	**60.62**	**42.36**	**48.27**	**41.82**	**41.29**
AWP	89.64	64.05	59.03	57.76	36.51	42.45	37.74	36.89	90.24	65.50	60.23	58.21	37.83	43.72	39.62	38.86
AWP_{HF}	**91.69**	**68.63**	**61.68**	**60.89**	**41.58**	**47.82**	**42.04**	**41.24**	**91.56**	**69.58**	**62.85**	**61.52**	**42.72**	**49.36**	**44.22**	**43.03**
HELP	90.34	62.65	56.05	53.87	36.23	40.67	34.97	34.64	91.34	67.65	57.32	54.93	37.26	42.15	38.71	37.89
$HELP_{HF}$	**91.79**	**66.04**	**59.60**	**57.77**	**42.86**	**45.52**	**41.78**	**41.23**	**92.79**	**69.70**	**61.64**	**58.64**	**43.38**	**46.85**	**43.34**	**41.83**

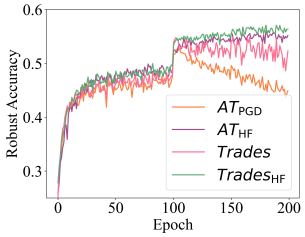

Figure 1. Learning curve.

		AT_{PGD}	AT_{HF}	Trades	$Trades_{HF}$
VGG16	Natural	80.97 ± 0.44	$\mathbf{82.04 \pm 0.37}$	81.43 ± 0.38	$\mathbf{82.38 \pm 0.29}$
	PGD_{50}	49.65 ± 0.17	$\mathbf{52.55 \pm 0.19}$	52.74 ± 0.23	$\mathbf{54.07 \pm 0.19}$
	AA	44.38 ± 0.06	$\mathbf{46.20 \pm 0.09}$	46.62 ± 0.07	$\mathbf{47.38 \pm 0.05}$
Moblie-NetV2	Natural	81.91 ± 0.21	$\mathbf{82.32 \pm 0.18}$	81.66 ± 0.31	$\mathbf{82.59 \pm 0.26}$
	PGD_{50}	51.00 ± 0.24	$\mathbf{52.68 \pm 0.20}$	53.30 ± 0.28	$\mathbf{54.43 \pm 0.34}$
	AA	46.47 ± 0.10	$\mathbf{47.05 \pm 0.08}$	47.30 ± 0.09	$\mathbf{47.91 \pm 0.06}$

Table 3. Results on CIFAR10 under VGG16 and MobileNetV2.

	FAT[3]	FAT_{HF}	Coreset-AT[2]	$Coreset\text{-}AT_{HF}$	N-FGSM[1]	$N\text{-}FGSM_{HF}$
Natural	83.22 ± 0.34	$\mathbf{83.84 \pm 0.38}$	80.64 ± 0.21	$\mathbf{80.76 \pm 0.18}$	80.39 ± 0.29	$\mathbf{81.03 \pm 0.32}$
$PGD_{50\text{-}10}$	46.12 ± 0.24	$\mathbf{48.44 \pm 0.19}$	45.44 ± 0.10	$\mathbf{46.46 \pm 0.11}$	48.03 ± 0.22	$\mathbf{48.92 \pm 0.12}$

Table 4. Performance in combination with other methods.

4. Performance on other networks

Given the pronounced disparities in the characteristics of different networks, we provide the results on CIFAR10 under VGG16 and MobileNetV2, respectively in Tab. 3. Each experiment is simulated for three times to avoid randomness. The results show that HFAT has better performance on different networks.

5. Combine HFAT with other methods

HFAT can be easily combined with other methods, since we can simply add additional gradient direction as momentum to the optimization direction of the standard AT. Fig. 4 studies the combined performance of HFAT with some methods that accelerate adversarial training. Each experiment is simulated for three times to avoid randomness. The superior performance not only further demonstrates the out-of-the-box usability of HFAT, but also provides us with ideas for accelerating HFAT.

References

[1] Pau de Jorge Aranda, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip Torr, Grégory Rogez, and Puneet Dokania. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35:12881–12893, 2022. 2

[2] Hadi M Dolatabadi, Sarah M Erfani, and Christopher Leckie. Adversarial coreset selection for efficient robust training. *International Journal of Computer Vision*, 131(12):3307–3331, 2023. 2

[3] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 2