Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation

Bo Huang[1,2], Mingyang Chen[1,2], Yi Wang[3], Junda Lu[4], Minhao Cheng[2], Wei Wang[*,1,2]

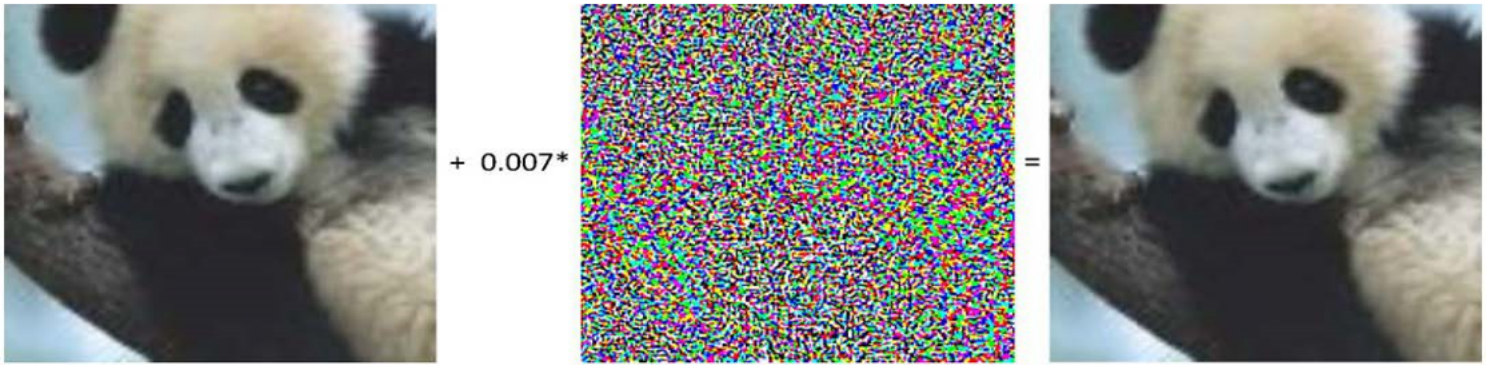
[1]DSA Thrust, Information Hub, HKUST (GZ), China
[2]HKUST, China
[3]Dongguan University of Technology, China
[4]Macquarie University, Australia

2023.06.22

Session: THU-PM-389

Introduction

- Teacher-student architectures [1], [2]
  - A means of computational-effective deployment in real-time applications and edge devices
- Adversarial attacks are constantly emerging [3]–[5].



Panda Adversarial Perturbation Gibbon

Challenge

There is a higher risk of student models to encounter adversarial attacks at the edge.

Adversarial Training

- Adversarial training (AT) has been studied and demonstrated effective in improving adversarial robustness.
  - Requiring over-parameterized models with large capacity
  - Leading to a severe trade-off between accuracy and robustness

Limitation

The gained robustness of student models with small capacity by adversarial training is not satisfactory.

Adversarial Distillation

- Enabling the student model to inherit not only the prediction accuracy but also the adversarial robustness from a robust teacher model.
- Existing methods mainly utilize fixed supervision targets to guide the distillation optimization process.
  - Imposing an overcorrection towards model smoothness, leading to the adversarial trade-off.
  - The student model does not fully interact with the teacher model, limiting the distillation performance.

Motivation

- Maximum point-to-point alignment between student and teacher models on full input distribution along with adversarial spaces.

$$\mathcal{L}_{\mathrm{AD}} = \iint \mathcal{D}(S(x+\delta), T(x+\delta))\, d\delta\, dx$$

Challenge

However, it is extraordinarily challenging to directly optimize $\mathcal{L}_{\mathrm{AD}}$ in practice.

The Proposed AdaAD

A min-max framework to approximately derive maximum point-to-point alignment between student and teacher models

$$\min \max_{\|\delta\|_p \leq \epsilon} \mathcal{D}(S(x+\delta), T(x+\delta)).$$

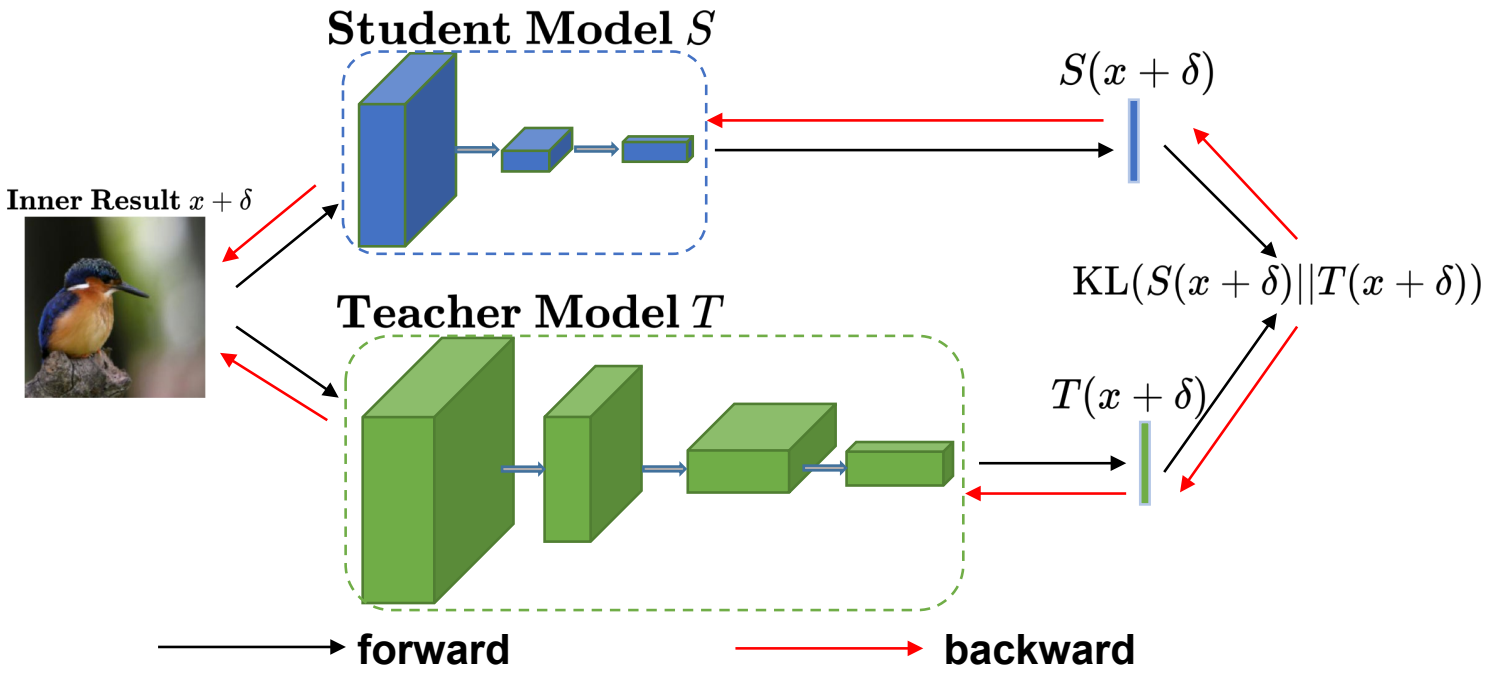
Inner Optimization

$$x^* = x + \arg\max_{\|\delta\|_p \leq \epsilon} \mathrm{KL}\left(S(x+\delta) \,||\, T(x+\delta)\right).$$

Outer Optimization

$$\arg\min_{\theta_S} \mathrm{KL}(S(x^*) \,||\, T(x^*)).$$

The Inner Optimization of AdaAD



The Benefits of AdaAD

- Reconciling accuracy and robustness
  - Existing methods inevitably impose an inductive bias towards local invariance [6], [7]:

$$\forall x \in \mathcal{B}(x_0, \epsilon), S(x) = p(y|x_0) \text{ or } S(x) = T(x_0)$$

  - AdaAD largely eliminates local invariance:

$$\forall x \in \mathcal{B}(x_0, \epsilon), S(x) = T(x)$$

- Allowing larger search radius ϵ
  - The inner results of AdaAD are not necessarily adversarial
  - Enabling the student model to inherit as much of accuracy and adversarial from the robust teacher model

Evaluation under White-box Attacks

Table: Recognition accuracy (%) under white-box attacks over CIFAR-10 dataset.

Teacher Model		WRN-34-20 [8]				
Model	Method	Clean	FGSM	PGD	CW_2	AA
RN-18	PGD-AT	82.95	57.16	52.87	77.56	47.69
	TRADES	83.00	58.42	53.18	76.92	49.21
	ARD	84.03	58.16	53.11	79.13	48.07
	IAD	84.71	61.28	54.92	79.44	49.85
	RSLAD	83.52	58.36	53.46	78.36	48.66
	AKD	83.22	58.63	54.16	78.44	49.26
	AdaAD	**85.58**	60.85	56.40	80.83	51.37
	AdaIAD	85.04	**62.62**	**58.34**	**81.15**	**52.96**
MN-V2	PGD-AT	77.54	53.58	49.90	72.54	44.56
	TRADES	79.80	54.84	50.51	75.30	45.67
	ARD	79.56	53.17	49.06	74.51	44.04
	IAD	83.31	58.29	52.98	78.03	47.11
	RSLAD	81.11	56.39	51.66	76.20	46.75
	AKD	83.41	57.71	52.35	77.97	46.82
	AdaAD	83.79	57.29	53.04	79.24	47.66
	AdaIAD	**84.63**	**59.79**	**54.97**	**80.21**	**49.29**

Significant gains of both clean and robust accuracy

Larger Search Radius ϵ

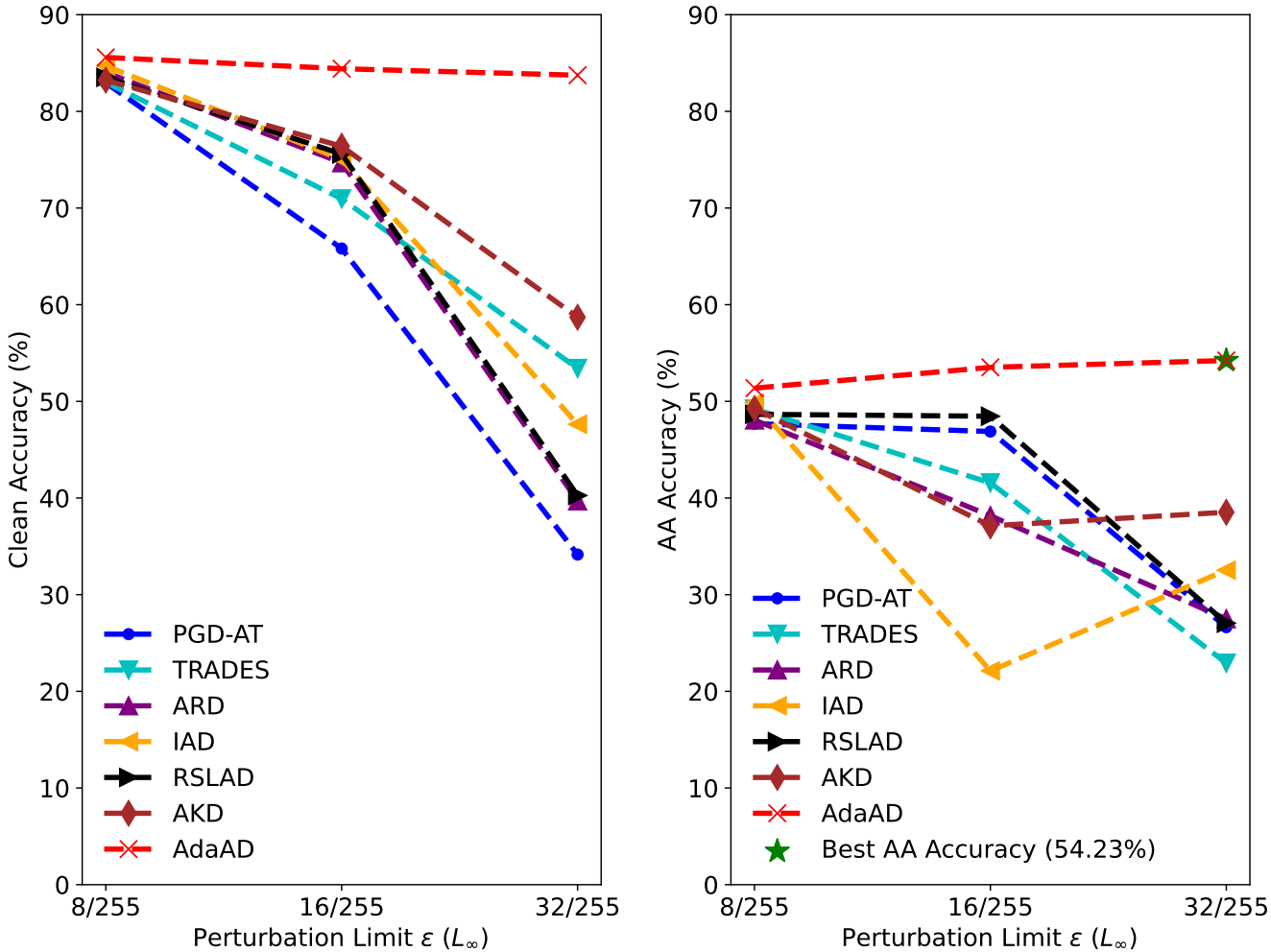


Figure: Clean and AA accuracy (%) of the studnet model ResNet-18 trained with increasing ϵ in the inner optimization on CIFAR-10.

Conclusion

- We formulate a new adversarial distillation objective by maximizing the prediction discrepancy between teacher and student models in the min-max framework.
- We design an adaptive adversarial distillation scheme, namely AdaAD, that adaptively searches for optimal *match points* in the inner optimization.
- Extensive experimental results verify that the performance of our method is significantly superior to that of the state-of-the-arts AT and AD methods in most scenarios.

References

[1] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. arXiv: 1503.02531.

[2] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman, "Teacher-student curriculum learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 9, pp. 3732–3740, 2020. DOI: 10.1109/TNNLS.2019.2934906.

[3] B. Biggio, I. Corona, D. Maiorca, *et al.*, "Evasion attacks against machine learning at test time," in *ECML PKDD*, 2013, pp. 387–402.

[4] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *ICLR*, 2014.

References

[5] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *ICML*, 2020, pp. 2206–2216.

[6] D. Stutz, M. Hein, and B. Schiele, "Confidence-calibrated adversarial training: Generalizing to unseen attacks," in *ICML*, 2020, pp. 9155–9166.

[7] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang, "Robust overfitting may be mitigated by properly learned smoothening," in *ICLR*, 2021.

[8] E. Chen and C. Lee, "LTD: low temperature distillation for robust adversarial training," *CoRR*, vol. abs/2111.02331, 2021. arXiv: 2111.02331.

[9] D. Wu, S. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *NeurIPS*, 2020, pp. 2958–2969.