Revisiting Adversarial Robustness Distillation from the Perspective of Robust Fairness

Xinli Yue Ningping Mou Qian Wang Lingchen Zhao

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University

Overview

- We pioneer research on robust fairness in adversarial robustness distillation (ARD), revealing incomplete fairness inheritance from teacher to student models and uncovering the internal reasons.

- We develop Fair-ARD, a framework enhancing robust fairness in knowledge transfer, adaptable to existing ARD techniques.

- Extensive experiments show Fair-ARD's superiority in improving student models' robust fairness and overall robustness.

Adversarial Robustness Distillation (ARD)

ARD aims to transfer the robustness of large teacher models to small student models, facilitating the attainment of robust performance on resource-limited devices.

$$\mathcal{L}_{\mathrm{ARD}} = (1-\alpha)\mathrm{CE}\left(S^{\tau}\left(x_i^j\right), y_i\right) + \alpha\tau^2\mathrm{KL}\left(S^{\tau}\left(\tilde{x}_i^j\right), T^{\tau}\left(x_i^j\right)\right)$$

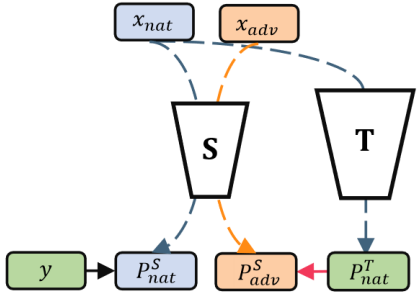
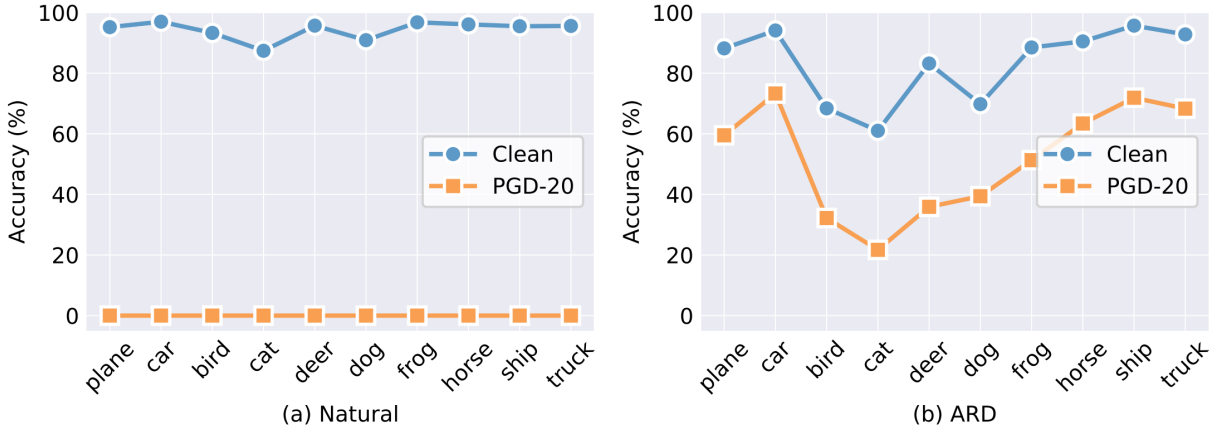


Figure from (Zi et al., 2021)

- Goldblum et al., Adversarially robust distillation, AAAI 2020
- Zi et al., Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better, ICCV 2021

Robust Fairness

Models trained by ARD may exhibit high robustness on some classes while demonstrating significantly low robustness on others.



- Xu et al., To be Robust or to be Fair: Towards Fairness in Adversarial Training, ICML 2021.

Problem Statement

The Challenge in ARD

- Traditional ARD often overlooks robust fairness, which often leads to uneven robustness in student models.

Research Focus

- Investigating whether ARD can effectively transfer robust fairness from large teacher models to smaller student models.
- Addressing the gaps in class-wise robustness.

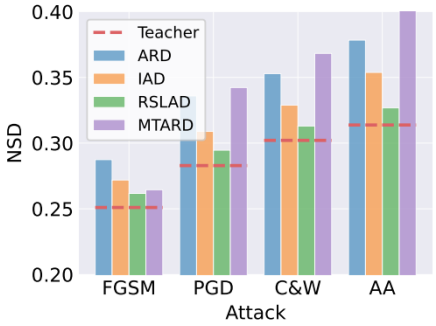
Objective

- To propose a method that addresses the robust fairness issue in ARD.

Findings on Robust Fairness Inheritance

Inheritance of Robust Fairness

- The robust fairness of the student models is worse than the teacher model.
- Student models can only partially inherit the teacher model's robust fairness.



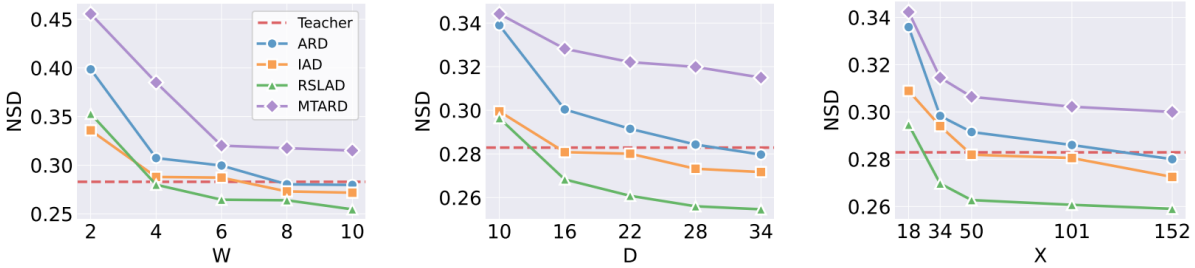
(a) ResNet18

- NSD: Normalized Standard Deviation of Class-wise Robustness.

Findings on Robust Fairness Inheritance

Impact of Model Capacity Gap

- The gap in capacity between teacher and student models affects the inheritance of robust fairness.



(b) WRN-34-W (c) WRN-D-10 (d) ResNet-X

Findings on Robust Fairness Inheritance

Disparity in Class Robustness Inheritance

- Different classes inherit different proportions of robustness from the teacher model.
- We should assign larger (smaller) weights to harder (easier) classes, allowing student models to focus more on the teacher model's knowledge about hard classes.

Method	FGSM		PGD		C&W		AA	
	cat	car	cat	car	cat	car	cat	car
ARD	85.15%	**94.95%**	80.60%	**92.78%**	80.74%	**93.59%**	72.53%	**93.13%**
IAD	84.55%	**96.39%**	81.72%	**94.42%**	86.89%	**95.26%**	79.83%	**94.69%**
RSLAD	90.30%	**95.07%**	88.81%	**93.79%**	92.62%	**94.36%**	89.70%	**94.56%**
MTARD	89.70%	**99.28%**	75.37%	**96.70%**	71.31%	**96.41%**	63.52%	**96.11%**

Fair-ARD

Overview of Fair-ARD

- A novel framework to address the issue of partial robust fairness inheritance in ARD.
- To ensure a more equitable distribution of class-wise robustness in the student model.

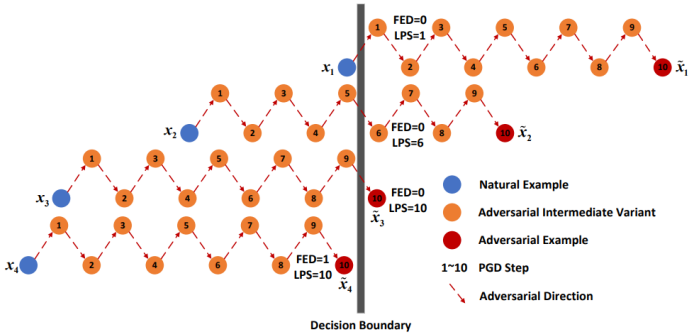
$$\min_{\theta_S} \frac{1}{C} \sum_{i=1}^{C} \frac{1}{n_i} \sum_{j=1}^{n_i} \omega_i \mathcal{L}_{\mathrm{ARD}} \left(S, T, x_i^j, y_i, \tau, \alpha \right)$$

Fair-ARD

Metrics to Measure Class Difficulty

- Compared with FED, LPS provides a more fine-grained example difficulty measure.

$$\begin{cases} \tilde{x}^{(t+1)} = \Pi_{\mathcal{B}_\epsilon[x]}\left(\tilde{x}^{(t)} + \gamma \mathrm{sign}\left(\nabla_{\tilde{x}^{(t)}} \ell\left(S\left(\tilde{x}^{(t)}\right), y\right)\right)\right) \\ d(x, y) = \underset{t \in [0,K]}{\mathrm{argmin}}\left(S\left(\tilde{x}^{(t)}\right) \neq y\right), \end{cases}$$



Fair-ARD

The Re-weighting Strategy

- Following the principle of giving larger (smaller) weights to harder (easier) classes, the weight ω_i should decrease w.r.t. κ_i.

$$\kappa_i = \frac{1}{n_i} \sum_{j=1}^{n_i} d\left(x_i^j, y_i\right)$$

$$\omega_i = \frac{1}{\kappa_i^{\beta}}$$

Evaluations

Worst-class Robustness

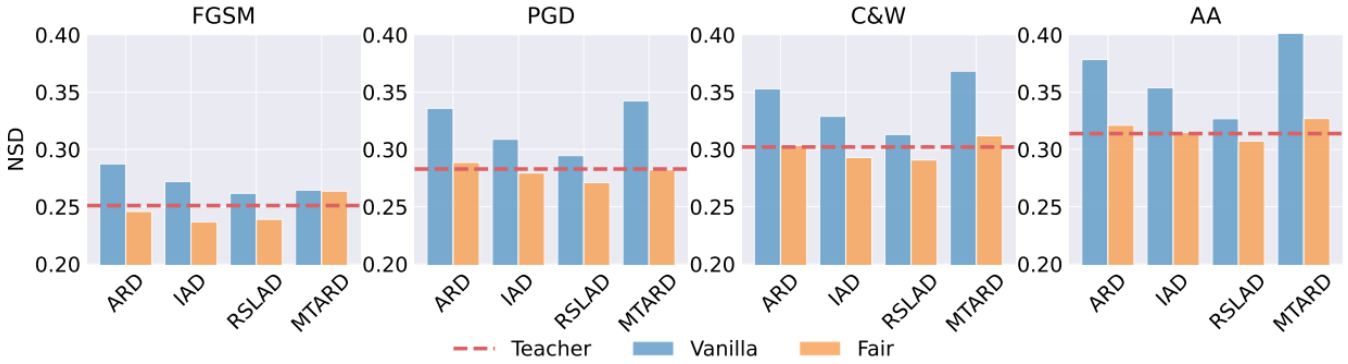
- Fair-ARD shows significant improvement in worst-class robustness under various adversarial attacks. This demonstrates the effectiveness of Fair-ARD in improving robust fairness.

Method	Clean		FGSM		PGD		C&W		AA	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
Natural	94.35	87.40	16.13	5.30	0.00	0.00	0.00	0.00	0.00	0.00
SAT	84.27	64.10	56.81	28.30	49.11	22.10	48.58	20.50	46.13	17.40
TRADES	82.22	64.80	58.38	31.00	52.35	25.80	50.33	22.30	49.01	21.20
ARD	**83.22**	61.00	**58.77**	28.10	51.65	21.60	**51.25**	19.70	49.05	16.90
Fair-ARD (ours)	82.96	**68.10**	57.69	**39.20**	**52.05**	**33.20**	50.69	**31.00**	**49.13**	**29.20**
IAD	**83.25**	60.60	**58.90**	27.90	52.08	21.90	51.01	21.20	48.95	18.60
Fair-IAD (ours)	83.19	**68.70**	58.31	**36.50**	**52.27**	**30.00**	**51.16**	**28.10**	**49.28**	**25.60**
RSLAD	83.04	62.70	60.03	29.80	54.13	23.80	52.76	22.60	51.18	20.90
Fair-RSLAD (ours)	**83.59**	**67.80**	**60.16**	**35.50**	**54.33**	**29.70**	**53.07**	**26.90**	**51.23**	**25.10**
MTARD	**87.14**	**69.80**	**60.62**	30.10	50.81	20.70	48.85	18.00	46.10	16.10
Fair-MTARD (ours)	81.98	62.10	59.11	**30.60**	**53.96**	**27.50**	**52.32**	**24.20**	**50.60**	**22.60**

Evaluations

NSD

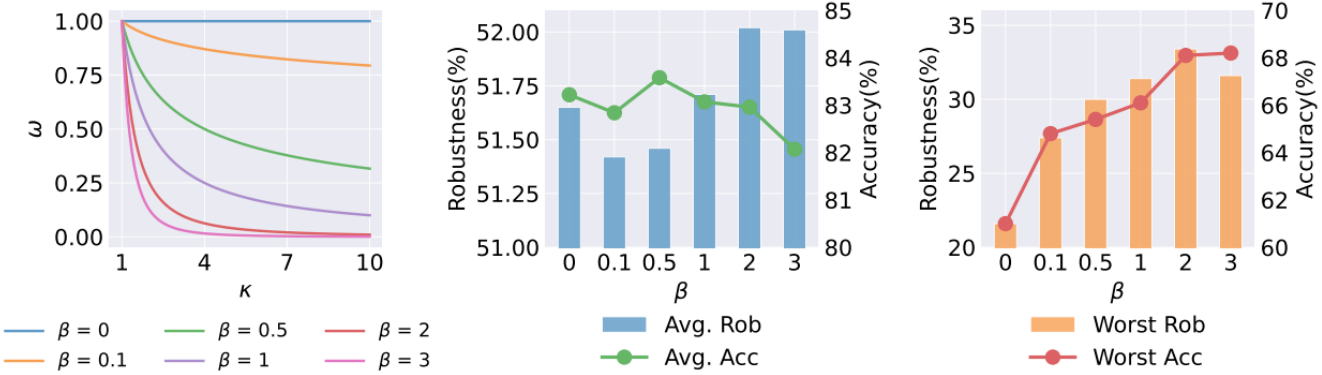
- Fair-ARD achieves lower NSD values, showing its effectiveness in balancing class-wise robustness.



Ablation Studies

The Effect of the Hyperparameter β

- Fair-ARD maintains its effectiveness across a range of $\boldsymbol{\beta}$, highlighting its adaptability.



Discussion

Comparison with GAIRAT and Robust Fairness Algorithms

- Fair-ARD outperforms others in achieving better worst-class robustness while maintains comparable overall robustness.

Method	Clean		FGSM		PGD		C&W		AA	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
ARD	**83.22**	61.00	**58.77**	28.10	51.65	21.60	**51.25**	19.70	49.05	16.90
GAIR-ARD	81.09	65.80	54.04	34.60	50.00	32.60	42.69	20.70	40.88	18.90
FRL-ARD	81.47	61.40	56.90	34.30	50.24	28.60	49.76	26.80	47.84	23.90
FAT-ARD	82.98	65.20	57.62	38.00	51.11	31.00	49.99	29.20	48.22	25.90
CFA-ARD	82.00	62.30	57.68	30.60	**52.43**	25.00	51.23	22.30	**49.73**	20.00
Fair-ARD (ours)	82.96	**68.10**	57.69	**39.20**	52.05	**33.20**	50.69	**31.00**	49.13	**29.20**

Conclusion

- We explore the issue of robust fairness in ARD, revealing the partial inheritance of robust fairness by student models from teacher models.
- We develope Fair-ARD, a novel framework that employs a refined class difficulty metric and a re-weighting strategy.
- We show that Fair-ARD significantly enhances the robust fairness of student models while maintaining high overall robustness, paving the way for fairer lightweight deep learning models.

Thanks!


lczhaocs@whu.edu.cn


https://github.com/NISP-official/Fair-ARD