Enhanced Accuracy and Robustness via Multi-Teacher Adversarial Distillation

Shiji Zhao [1,2], Jie Yu [1,2], Zhenlong Sun [2], Bo Zhang [2], Xingxing Wei [1*]

1.Institute of Artificial Intelligence, Hangzhou Innovation Institute, Beihang University, Beijing, China

2.WeChat Search Application Department, Tencent, Beijing, China

TEL AVIV 2022

Problem Presentation and Contribution

Problem Presentation:

Adversarial training has several shortcomings in some general scenes. Firstly, the robustness of models obtained from adversarial training is related to the size of model. Secondly, the accuracy of identifying clean examples by adversarial trained DNNs is far worse than normal trained DNNs.

Our contributions:

- We propose a novel adversarial robustness distillation method called Multi-Teacher Adversarial Robustness Distillation (MTARD).
- We design a joint training algorithm based on the proposed Adaptive Normalization Loss to balance the influence on the student model between the adversarial teacher model and the clean teacher model.
- We empirically verify the effectiveness of MTARD in improving the performance of small models. For the models trained by our MTARD, the Weighted Robust Accuracy has been greatly improved compared with the Multi-Teacher Adversarial Robustness Distillation state-of-the-art adversarial training and distillation method against white-box and black-box attacks.

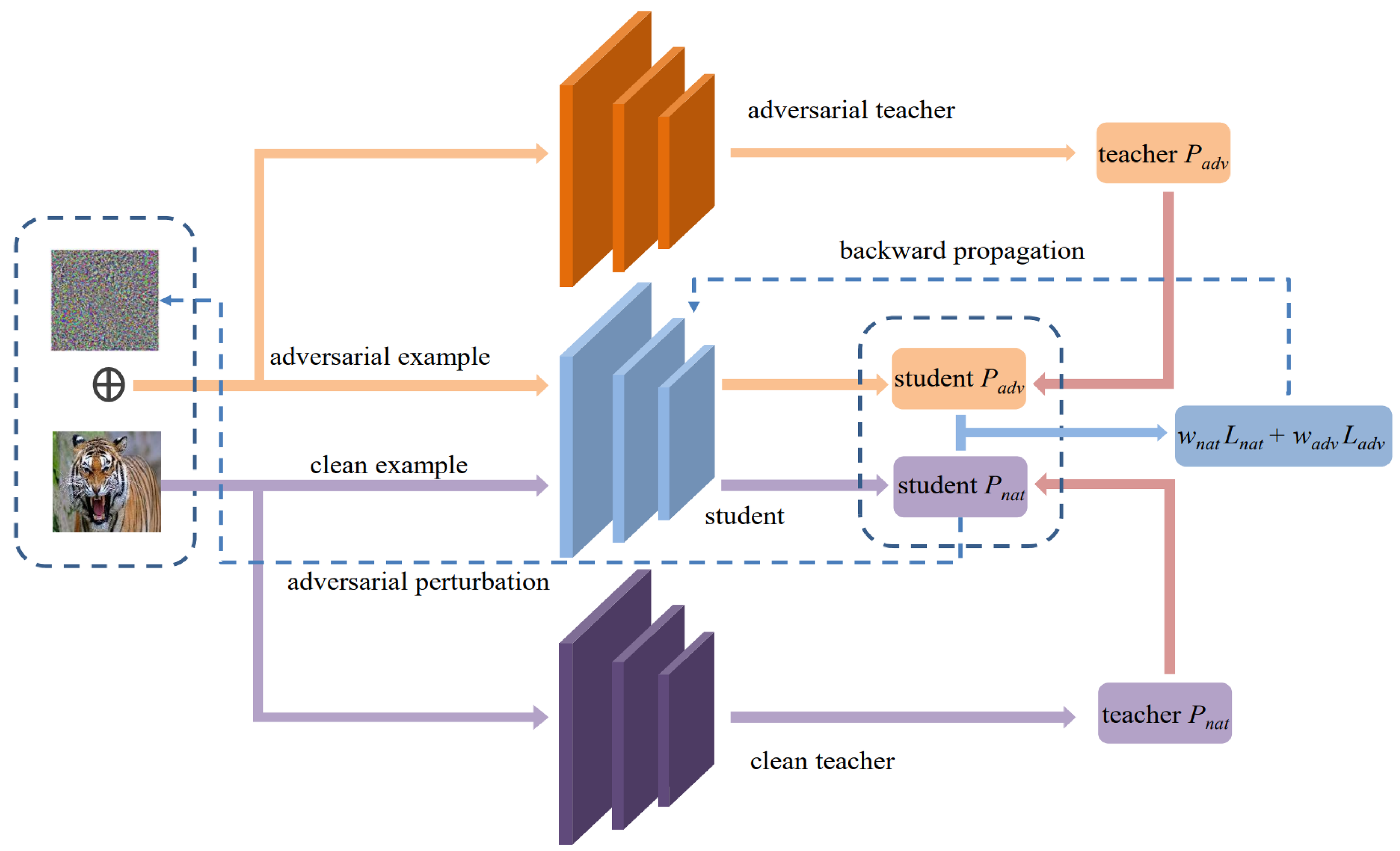
Multi-Teacher Adversarial Robustness Distillation

$$x_{adv} = \underset{\delta \in \Omega}{\arg\max}\, CE(S(x_{nat}+\delta;\theta_S), y)$$

$$\underset{\theta_S}{\arg\min}(1-\alpha)KL(S(x_{nat}), T_{nat}(x_{nat})) + \alpha KL(S(x_{adv}), T_{adv}(x_{adv}))$$

The student can learn robustness from the adversarial teacher and the ability to identify clean examples from the clean teacher. The inputs of the clean teacher are initial clean examples from original datasets. In contrast, the inputs of the adversarial teacher are adversarial examples produced by the student model. The student inputs are divided into clean examples and adversarial examples. The outputs of clean examples and adversarial examples will be guided by adversarial soft label and clean soft label.

Framework



Adaptive Normalization Loss In MTARD

In order to get both clean and robust accuracy, a strategy is needed to balance the influence between the adversarial teacher and the clean teacher.

$$L_{total}(t) = w_{adv}(t)L_{adv}(t) + w_{nat}(t)L_{nat}(t)$$

$$w_{adv}(t) = \frac{r_w[L_{adv}(t)/L_{adv}(0)]^{\beta}}{[L_{nat}(t)/L_{nat}(0)]^{\beta} + [L_{adv}(t)/L_{adv}(0)]^{\beta}} + (1-r_w)w_{adv}(t-1)$$

$$w_{nat}(t) = 1 - w_{adv}(t)$$

Adaptive Normalization Loss

Adaptive Normalization Loss used in MTARD can inhibit the rapid growth of a stronger teacher throughout the training cycle. Adaptive Normalization Loss can dynamically suppress the teacher's teaching ability by controlling the loss weight, while the ability of the other teacher will become stronger in the following period.

$$L_{total}(t) = \sum_{i=1}^{N} w_i(t)L_i(t)$$

$$\tilde{L}_i(t) = L_i(t)/L_i(0)$$

$$r_i(t) = [\tilde{L}_i(t)]^{\beta} / \sum_{i=1}^{N}[\tilde{L}_i(t)]^{\beta}$$

$$w_i(t) = r_w r_i(t) + (1-r_w)w_i(t-1)$$

Part Experimental Results

Robustness Evaluation

Table 2. White-box robustness of ResNet-18 on CIFAR-10 and CIFAR-100 dataset.

		CIFAR-10			CIFAR-100		
Attack	Defense	Clean	Robust	W-Robust	Clean	Robust	W-Robust
FGSM	Natural	94.57%	18.60%	56.59%	75.18%	7.96%	41.57%
	SAT	84.2%	55.59%	69.90%	56.16%	25.88%	41.02%
	TRADES	83.00%	58.35%	70.68%	57.75%	31.36%	44.56%
	ARD	84.11%	58.4%	71.26%	60.11%	33.61%	46.86%
	RSLAD	83.99%	60.41%	72.2%	58.25%	34.73%	46.49%
	MTARD	87.36%	61.2%	**74.28%**	64.3%	31.49%	**47.90%**
PGD_{sat}	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	TRADES	83.00%	52.35%	67.68%	57.75%	28.05%	42.9%
	SAT	84.2%	45.95%	65.08%	56.16%	21.18%	38.67%
	TRADES	83.00%	52.35%	67.68%	57.75%	28.05%	42.9%
	ARD	84.11%	50.93%	67.52%	60.11%	29.4%%	**44.76%**
	RSLAD	83.99%	53.94%	68.97%	58.25%	31.19%	44.72%
	MTARD	87.36%	50.73%	**69.05%**	64.3%	24.95%	44.63%
PGD_{trades}	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	SAT	84.2%	48.12%	66.16%	56.16%	22.02%	39.09%
	TRADES	83.00%	53.83%	68.42%	57.75%	28.88%	43.32%
	ARD	84.11%	52.96%	68.54%	60.11%	30.51%	45.31%
	RSLAD	83.99%	55.73%	69.86%	58.25%	32.05%	45.15%
	MTARD	87.36%	53.60%	**70.48%**	64.3%	26.75%	**45.53%**
CW_{∞}	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	SAT	84.2%	45.97%	65.09%	56.16%	20.9%	38.53%
	TRADES	83.00%	50.23%	66.62%	57.75%	24.19%	40.97%
	ARD	84.11%	50.15%	67.13%	60.11%	27.56%	43.84%
	RSLAD	83.99%	52.67%	**68.33%**	58.25%	28.21%	43.23%
	MTARD	87.36%	48.57%	67.97%	64.3%	23.42%	**43.86%**

Table 4. Black-box robustness of ResNet-18 on CIFAR-10 and CIFAR-100 dataset.

		CIFAR-10			CIFAR-100		
Attack	Defense	Clean	Robust	W-Robust	Clean	Robust	W-Robust
PGD-20	SAT	84.2%	64.74%	74.47%	56.16%	38.1%	47.13%
	TRADES	83.00%	63.56%	73.28%	57.75%	38.2%	47.98%
	ARD	84.11%	63.59%	73.85%	60.11%	39.53%	49.82%
	RSLAD	83.99%	63.9%	73.95%	58.25%	39.93%	49.09%
	MTARD	87.36%	65.17%	**76.27%**	64.3%	41.39%	**52.85%**
CW_{∞}	SAT	84.2%	64.88%	74.54%	56.16%	39.42%	47.79%
	TRADES	83.00%	62.85%	72.93%	57.75%	38.63%	48.19%
	ARD	84.11%	62.78%	73.45%	60.11%	38.85%	49.48%
	RSLAD	83.99%	63.02%	73.51%	58.25%	39.67%	48.96%
	MTARD	87.36%	64.65%	**76.01%**	64.3%	41.03%	**52.67%**
SA	SAT	84.2%	71.3%	77.75%	56.16%	41.27%	48.72%
	TRADES	83.00%	70.33%	76.67%	57.75%	41.96%	49.86%
	ARD	84.11%	73.3%	78.71%	60.11%	48.79%	54.45%
	RSLAD	83.99%	72.1%	78.05%	58.25%	45.34%	51.80%
	MTARD	87.36%	79.99%	**83.68%**	64.3%	41.03%	**52.67%**

Ablation Studies

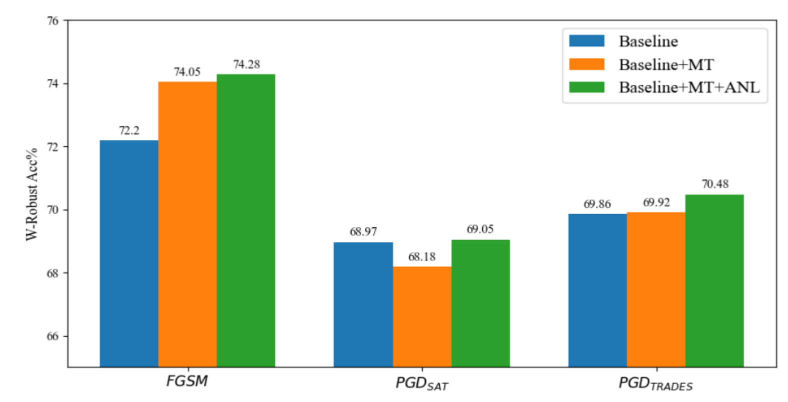


Fig. 2. Ablation study with ResNet-18 student network distilled using variants of our MTARD and Baseline method on CIFAR-10. MT and ANL are abbreviations of multi-teacher and Adaptive Normalization Loss. Baseline+MT means using multiple teachers in Baseline. Baseline+MT+ANL means our MTARD method.

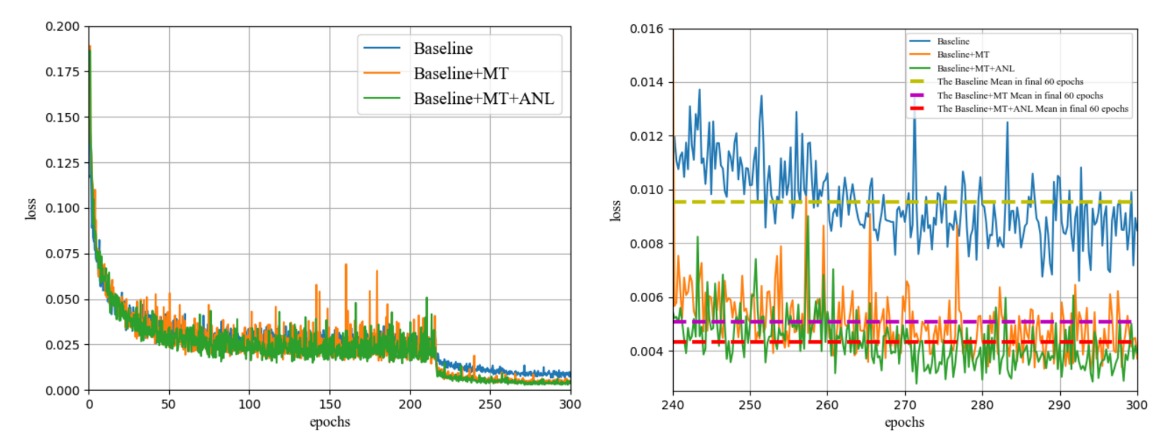


Fig. 3. The training loss with ResNet-18 student network distilled using variants of Baseline, Baseline+MT, and Baseline+MT+ANL (our MTARD) on CIFAR-10. MT and ANL are abbreviations of multi-teacher and Adaptive Normalization Loss. The y axis is the L_{total} in the training epoch x. The left is the change curve of L_{total} in the whole training process, the right is the change curve of L_{total} in final 60 epochs.