– Supplemental Material –
Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better

Bojia Zi[1,2]*, Shihao Zhao[1,2]*, Xingjun Ma[3]†, Yu-Gang Jiang[1,2]†
[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan Univeristy
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]School of Information Technology, Deakin University, Geelong, Australia

A. Hyper-parameter Selection

In this section, we explore the impact of the hyper-parameter α in equation (3). We apply RSLAD to train and distill ResNet-18 students from the the WideResNet-34-10 teacher on CIFAR-10 using different α. We show the robust accuracy against $\text{PGD}_{\text{TRADES}}$ with respect to $k = \frac{\alpha}{(1-\alpha)}$, which is the ratio of the adversarial loss term to the natural loss term. We report the robustness results at the best checkpoints in Figure 1. It can be observed that robustness rises rapidly with the increase of the ratio k and reaches a plateau after $k = 1.0$. When the ratio becomes larger than 1, the robustness fluctuates slightly around 55.9% and achieves the best at $k = 5.0/1.0$.

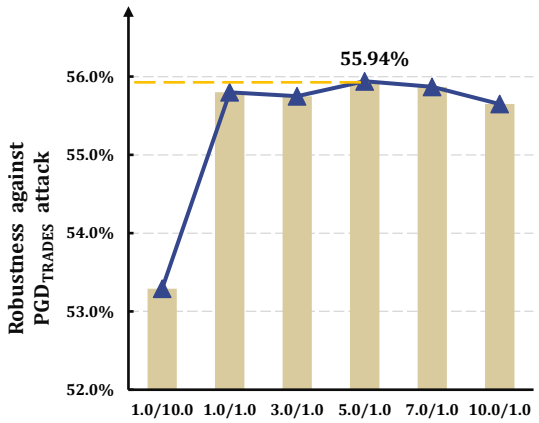

Figure 1: Robustness on CIFAR-10 for ResNet-18 student distilled by our RSLAD with WideResNet-34-10 teacher, under different hyper-parameters α. **X-axis:** ratio of the adversarial term to the natural term: $k = \frac{\alpha}{(1-\alpha)}$; **Y-axis:** robustness against $\text{PGD}_{\text{TRADES}}$ on CIFAR-10 test set.

B. Learning From Different Teachers

In Section 4.4, we have demonstrated how to choose a good teacher network and showed the impact of the teacher on the student's robustness. Here, we show a more complete robustness results of the student network against all 5 attacks mentioned in Section 4.1. We report results at both the best and the last checkpoints in Table 2. The teachers' robustness is shown in Table 1. We can confirm the phenomenon of *robust saturation* and *robust underfitting* according to more evaluation attacks. This indicates that a moderately large teacher network can be a better teacher than a overly large teacher network.

C. RSLs of Natural or Adversarial Examples?

RSLs are the outputs of a robust model, however, it can be on either the natural examples (natural RSLs $T(x)$) or the adversarial examples (adversarial RSLs $T(x')$). Same as TRADES, MART, ARD and IAD, RSLAD utilizes the $T(x)$ as RSLs. But one may wonder whether $T(x')$ is better than $T(x)$. To answer this question, we replace the $T(x)$ used in our RSLAD loss terms (Equation 3) with $T(x')$. This experiment is conducted with ResNet-18 student and WideResNet-34-10 teacher on CIFAR-10 dataset. We report the results at the best checkpoints in Table 3.

Looking at the first row of Table 3, one can find that, when replacing $T(x')$ with $T(x)$ in all loss terms, the robustness reaches 49.79% against AA, which is slightly higher than that of the TRADES (49.27%) (see Table 2), but still much lower than our RSLAD with $T(x)$. Moving on to the second and third rows, one may notice that, when replacing more $T(x')$ with $T(x)$ in $\mathcal{L}_{max}$ or $\mathcal{L}_{min}$,

*Equal contribution: Bojia Zi(bjzi19@fudan.edu.cn) and Shihao Zhao(shzhao19@fudan.edu.cn)
†Correspondence to Xingjun Ma (daniel.ma@deakin.edu.au) and Yu-Gang Jiang (ygj@fudan.edu.cn)

Table 1: Robustness of the teacher networks used in our experiments on CIFAR-10 dataset. The maximum perturbation is $\epsilon = 8/255$. The best results are **blodfaced**.

Teacher	Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_∞	AA
ResNet-18	81.93%	57.49%	52.66%	53.68%	50.58%	49.23%
ResNet-34	83.38%	59.54%	53.70%	55.25%	52.39%	50.46%
ResNet-50	84.25%	60.27%	53.71%	55.31%	52.47%	50.47%
WideResNet-34-10	84.92%	60.87%	55.33%	56.61%	53.98%	53.08%
WideResNet-34-20	**85.64%**	**64.29%**	**59.86%**	**60.82%**	58.04%	56.86%
WideResNet-70-16	85.29%	64.20%	59.66%	60.46%	**58.60%**	**57.20%**

Table 2: Robustness of ResNet-18 student distilled using our RSLAD with 6 different teacher networks. Both the best and last checkpoints are reported and the best results are **blodfaced**.

Student	Teacher	Best Checkpoint						Last Checkpoint					
		Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_∞	AA	Clean	FGSM	PGD_{SAT}	PGD_{TRADES}	CW_∞	AA
RN-18	RN-18	81.14%	58.62%	53.92%	55.31%	52.08%	50.75%	81.43%	58.61%	53.63%	54.90%	51.90%	50.65%
	RN-34	81.96%	59.00%	53.94%	55.19%	51.95%	50.64%	81.79%	59.12%	53.78%	55.10%	52.03%	50.61%
	RN-50	82.19%	59.39%	54.09%	55.55%	52.24%	50.80%	82.46%	59.45%	53.57%	55.26%	51.93%	50.29%
	WRN-34-10	**83.38%**	60.01%	**54.24%**	**55.94%**	**53.30%**	**51.49%**	**83.33%**	59.90%	54.14%	**55.61%**	**53.22%**	**51.32%**
	WRN-34-20	83.36%	**60.11%**	54.18%	55.85%	51.58%	50.49%	83.25%	**60.31%**	**54.24%**	55.60%	51.86%	50.50%
	WRN-70-16	82.96%	59.61%	53.63%	55.17%	51.82%	50.27%	82.99%	59.42%	53.36%	54.96%	51.64%	50.04%

both clean accuracy and robustness can be improved. This clearly demonstrates the advantage of using RSLs of the natural examples, albeit RSLs of adversarial examples also help robustness.

D. Training all baselines for 300 epochs

Whilst the training epoch is 100 in SAT, TRADES and 200 in ARD, IAD, we train our RSLAD models for 300 epochs. For a fair comparison, here we also run the baseline methods for 300 epochs. Table 4 shows the results of 300-epoch SAT, TRADES, ARD and IAD on CIFAR-10, following the setting in Section 4.1 It shows that, although the robustness of **best** checkpoints has been slightly improved when training for 300 epochs using TRADES, ARD and IAD, their performances on the **last** checkpoints degrade, resulting in more overfitting. As expected, our RSLAD method achieves the best overall performance.

E. Teacher models' diversity

Considering that we use a teacher trained by TRADES in most experiments in the main text, to figure out whether our method works with different kinds of teachers, we try to train the teacher model using SAT+AWP. AWP [**?**] is used to boost the teacher's robustness and brings a robust accuracy of 54% against AA. The results are illustrated in Table 5. It shows that the **best** checkpoint of student model has 0.49% and 0.13% improvement in natural and robust accuracy, respectively, and the **last** checkpoint slightly degrades in robustness but gains 0.7% improvement in natural accuracy. It can be concluded that RSLAD can indeed boost the small models' robustness with various kinds teacher models.

Table 3: Robustness of ResNet-18 student trained using different types of RSLs (x: clean examples; x': adversarial examples) on CIFAR-10 dataset. The maximum perturbation is $\epsilon = 8/255$. The best results are **blodfaced**.

$\mathcal{L}_{min}$	$\mathcal{L}_{max}$	Clean	FGSM	PGD_{SAT}	$\text{PGD}_{\text{TRADES}}$	CW_∞	AA
$(1-\alpha)\text{KL}(S(x), \mathbf{T}(\mathbf{x}')) + \alpha\text{KL}(S(x'), \mathbf{T}(\mathbf{x}'))$	$\text{KL}(S(x'), \mathbf{T}(\mathbf{x}'))$	78.86%	57.16%	52.81%	54.12%	51.34%	49.79%
$(1-\alpha)\text{KL}(S(x), T(x)) + \alpha\text{KL}(S(x'), \mathbf{T}(\mathbf{x}'))$	$\text{KL}(S(x'), \mathbf{T}(\mathbf{x}'))$	79.04%	57.53%	53.14%	54.41%	51.37%	49.83%
$(1-\alpha)\text{KL}(S(x), T(x)) + \alpha\text{KL}(S(x'), \mathbf{T}(\mathbf{x}'))$	$\text{KL}(S(x'), T(x))$	82.95%	59.81%	54.13%	55.91%	53.06%	51.26%
$(1-\alpha)\text{KL}(S(x), T(x)) + \alpha\text{KL}(S(x'), T(x))$	$\text{KL}(S(x'), T(x))$	**83.38%**	**60.01%**	**54.24%**	**55.94%**	**53.30%**	**51.49%**

Table 4: Robustness results of ResNet-18 on CIFAR-10. The best results are **boldfaced**. -300 means 300 epochs of training.

Method	**Best Checkpoint**						**Last Checkpoint**					
	Clean	FGSM	PGD_{SAT}	$\text{PGD}_{\text{TRADES}}$	CW_∞	AA	Clean	FGSM	PGD_{SAT}	$\text{PGD}_{\text{TRADES}}$	CW_∞	AA
Natural	**94.65%**	0.0%	0.0%	0.0%	0.0%	0.0%	**94.65%**	0.0%	0.0%	0.0%	0.0%	0.0%
SAT	83.38%	56.41%	49.11%	51.11%	48.67%	45.83%	84.44%	55.37%	46.22%	48.72%	47.14%	43.64%
SAT-300	83.96%	55.42%	47.04%	49.21%	47.47%	44.73%	84.29%	52.08%	42.42%	44.69%	48.33%	40.99%
TRADES	81.93%	57.49%	52.66%	53.68%	50.58%	49.23%	82.20%	57.86%	52.30%	53.66%	50.69%	49.27%
TRADES-300	82.06%	57.97%	52.65%	53.96%	50.91%	49.50%	82.79%	57.50%	49.97%	51.83%	49.51%	47.59%
ARD	83.93%	59.31%	52.05%	54.20%	51.22%	49.19%	84.23%	59.33%	51.52%	53.74%	51.24%	48.90%
ARD-300	84.40%	59.81%	52.36%	54.49%	51.58%	49.70%	85.01%	55.42%	51.59%	53.59%	50.98%	48.72%
IAD	83.24%	58.60%	52.21%	54.18%	51.25%	49.10%	83.90%	58.95%	51.35%	53.15%	50.52%	48.48%
IAD-300	83.68%	59.20%	52.83%	54.58%	51.84%	49.54%	84. 35%	59.92%	51.30%	53.44%	50.61%	48.60%
RSLAD	83.38%	**60.01%**	**54.24%**	**55.94%**	**53.30%**	**51.49%**	83.33%	**59.90%**	**54.14%**	**55.61%**	**53.22%**	**51.32%**

Table 5: White-box robustness results of ResNet-18 on CIFAR-10. The best results are **boldfaced**. * indicates the teacher is SAT+AWP.

Method	**Best Checkpoint**			**Last Checkpoint**		
	Clean	$\text{PGD}_{\text{TRADES}}$	AA	Clean	$\text{PGD}_{\text{TRADES}}$	AA
Natural	94.65%	0.0%	0.0%	94.65%	0.0%	0.0%
SAT	83.38%	51.11%	45.83%	84.44%	48.72%	43.64%
TRADES	81.93%	53.68%	49.23%	82.20%	53.66%	49.27%
RSLAD	83.38%	55.94%	51.49%	83.33%	55.61%	**51.32%**
RSLAD*	83.87%	**56.78%**	**51.62%**	84.03%	**56.52%**	51.09%