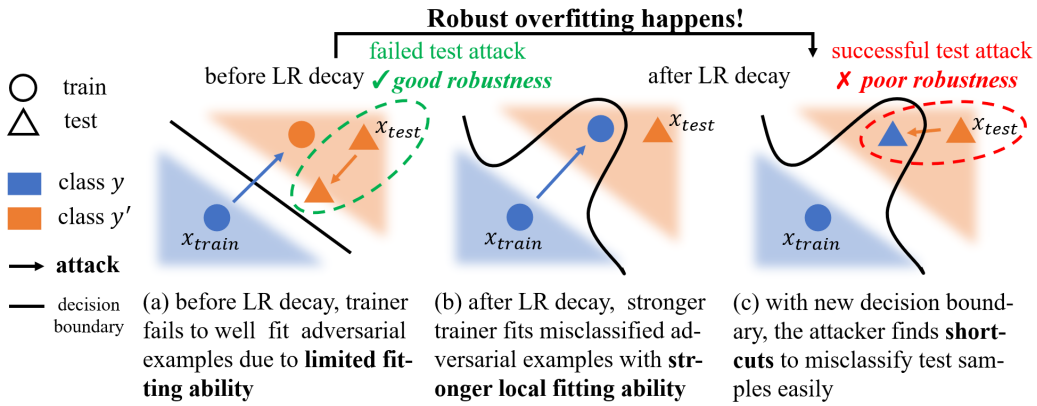

Figure A: An illustration of the rise of robust overfitting with a 2-dimensional example. Here, the attack (from the clean example to the adversarial example) follows the direction to the nearest decision boundary.

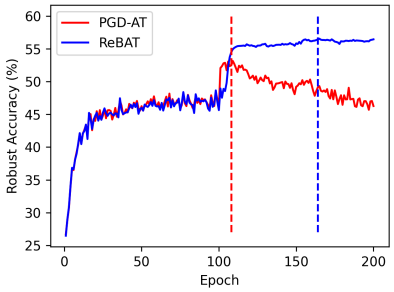

Figure B: Test robust accuracy of PGD-AT and ReBAT on CIFAR-10 under the perturbation norm $\varepsilon_\infty = 8/255$ based on the PreActResNet-18 architecture.

Table A: Performance on CIFAR-10 with different backbone networks.

Method	PreActResNet-18		WideResNet-34-10	
	best AA	final AA	best AA	final AA
AWP	50.09	49.85	53.32	53.38
MLCAT$_{WP}$	50.70	50.32	54.65	54.56
ReBAT	51.13	51.22	54.78	54.80
ReBAT[strong]	**51.49**	**51.39**	54.80	54.91
ReBAT+CutMix	50.18	50.22	**55.75**	**55.72**

Table B: Performance comparison on CIFAR-10 with PreActResNet-18 backbone.

Method	Natural		AutoAttack	
	best	final	best	final
AT (our reproduction)	81.61	**84.67**	47.51	42.31
SAAT (original paper)	76.37	76.38	48.86	47.17
AWP (our reproduction)	81.11	80.62	50.09	49.85
AWP+SAAT (original paper)	79.49	77.91	50.67	49.29
Ours: stronger attack	78.17	80.25	50.99	47.66
Ours: ReBAT (BoAT loss + small decay factor)	**81.86**	81.91	51.13	51.22
Ours: ReBAT[strong] (ReBAT + stronger attack)	78.71	78.85	**51.49**	**51.39**

Table C: Performance comparison on CIFAR-10 with PreActResNet-18 backbone.

Method	Natural		AutoAttack	
	best	final	best	final
AT (baseline)	81.61	84.67	47.51	42.31
AT+WA (baseline)	83.50	84.94	49.89	43.83
AWP (baseline)	81.11	80.62	50.09	49.85
BoAT loss	81.54	82.42	50.56	48.98
Small decay factor	81.90	82.39	50.81	50.58
ReBAT (BoAT loss + small decay factor)	81.86	81.91	51.13	51.22
stronger attack	78.17	80.25	50.99	47.66
ReBAT[strong] (BoAT loss + small decay factor + stronger attack)	78.71	78.85	51.49	51.39