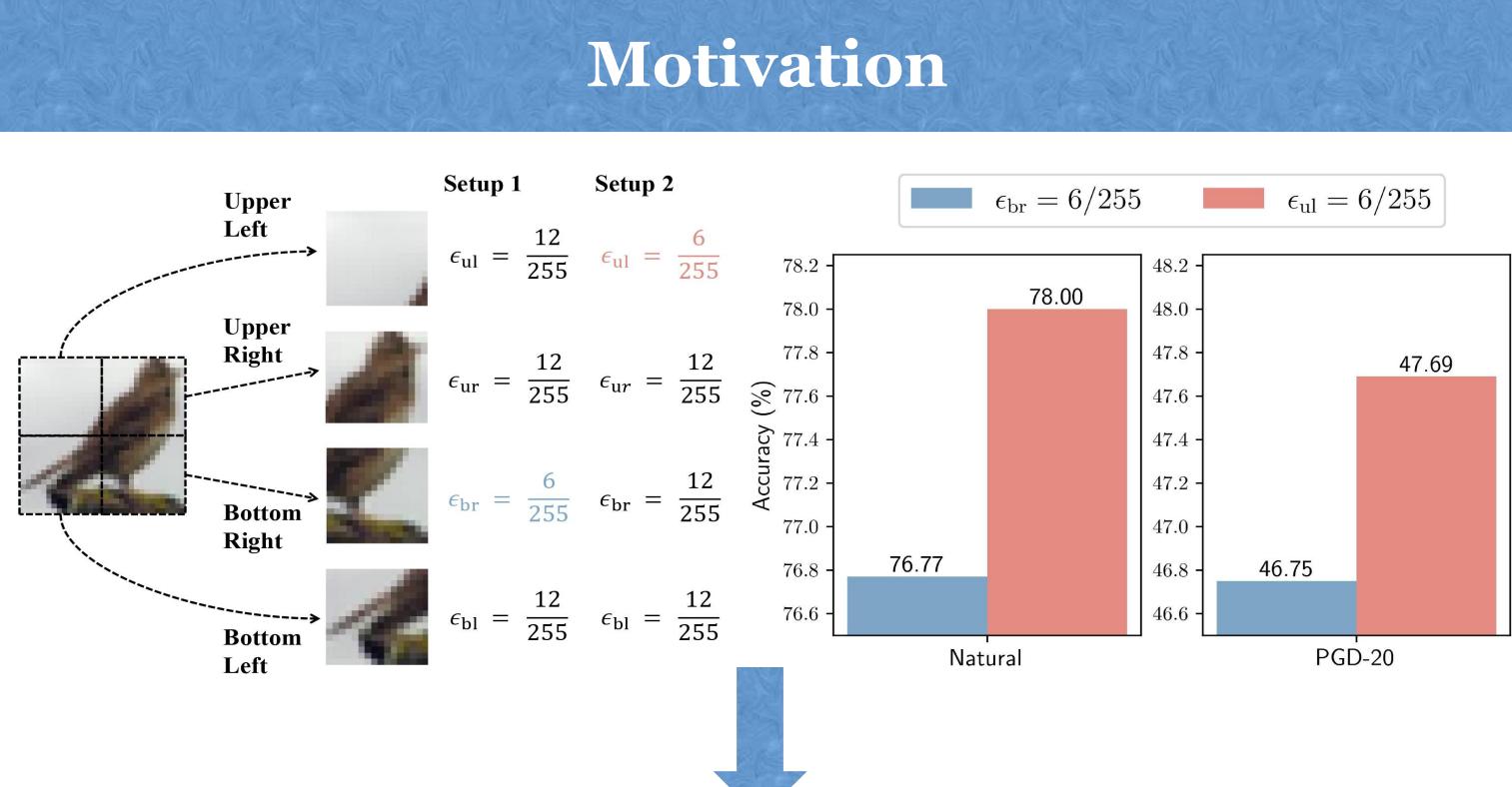
Improving Accuracy-robustness Trade-off via Pixel Reweighted Adversarial Training TMLR De ILCOL

Jiacheng Zhang, Feng Liu*, Dawei Zhou, Jingfeng Zhang, Tongliang Liu* iiachengzhang.ml@gmail.com THE UNIVERSITY OF MELBOURNE THE UNIVERSITY OF MELBOURNE THE UNIVERSITY OF MELBOURNE jiachengzhang.ml@gmail.com

Background

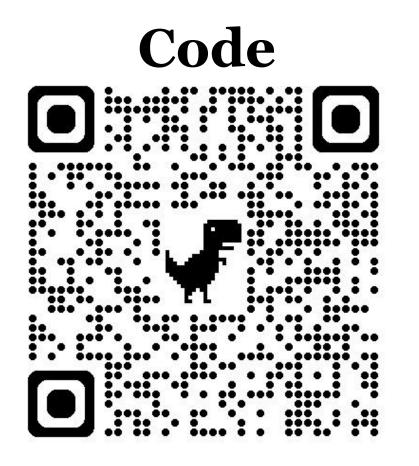
- > Adversarial training (AT) trains models using adversarial examples (AEs), which are natural images modified with specific perturbations to mislead the model.
- > These perturbations are constrained by a predefined perturbation budget ϵ and are *equally* applied to each pixel within an image.



Changing the perturbation budgets for different parts of an image has the potential to boost robustness and accuracy at the same time. Different pixel regions contribute *differently* to robustness and accuracy.

Pixels contribute more towards higher robustness and accuracy should have more weight.







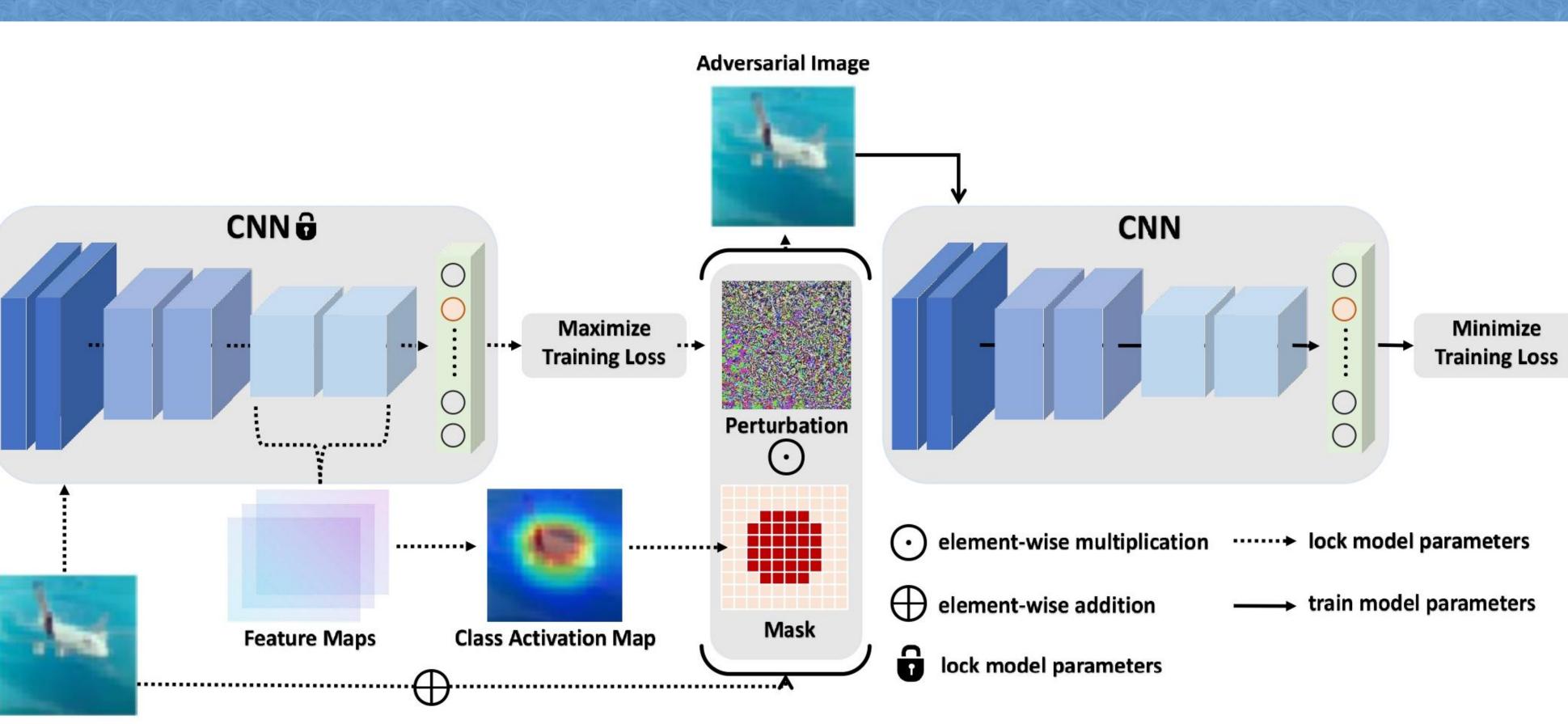






Main Insight Guiding the model to focus more on essential pixel regions during training can help improve the generalizability of vision models.

Pixel-reweighted AdversaRial Training (PART)



Natural Image

- > PART leverages the power of *Class Activation Mapping* (CAM) methods to identify important pixel regions.
- \triangleright PART partially reduces ϵ for less influential pixels, guiding the model to focus more on key regions that affect its outputs.
- > The innovation in the generation process of AEs allows PART to be orthogonal to many AT methods (e.g., TRADES, MART), and thus PART can be easily integrated into existing AT methods.
- > PART-based methods align better with semantic information (see results).
- > In general, PART serves as a general idea rather than a specific method, and CAM is used as one of the tools to realize the idea.

What's Next?

- \triangleright Design better algorithms to reweight pixels.
- \triangleright Extend the work to Transformers (e.g., leveraging the attention mechanism).

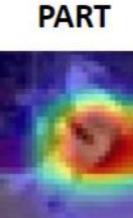












(a) Epoch



PART



Dataset	Method	Natural	PGD-20	MMA	AA
		Ι	ResNet-18		
CIFAR-10	AT PART $(s = 1)$ PART $(s = 10)$	$\begin{array}{r} 82.58 \pm 0.14 \\ 83.42 \pm 0.26 \ (+ \ 0.84) \\ \textbf{83.77} \pm \textbf{0.15} \ (+ \ \textbf{1.19}) \end{array}$	$\begin{array}{r} \textbf{43.69} \pm \textbf{0.28} \\ \textbf{43.65} \pm \textbf{0.16} (-\textbf{ 0.04}) \\ \textbf{43.36} \pm \textbf{0.21} (-\textbf{ 0.33}) \end{array}$	$\begin{array}{c} 41.80 \pm 0.10 \\ \textbf{41.98} \pm \textbf{0.03} \ \textbf{(+0.18)} \\ 41.83 \pm 0.07 \ \textbf{(+0.03)} \end{array}$	$\begin{array}{c} 41.63 \pm 0.22 \\ \textbf{41.74} \pm \textbf{0.04} \ (\textbf{+ 0.11}) \\ 41.41 \pm 0.14 \ (\textbf{- 0.22}) \end{array}$
	TRADES PART-T $(s = 1)$ PART-T $(s = 10)$	78.16 ± 0.15 $79.36 \pm 0.31 (+ 1.20)$ $80.13 \pm 0.16 (+ 1.97)$	$\begin{array}{c} 48.28 \pm 0.05 \\ \textbf{48.90} \pm \textbf{0.14} \ \textbf{(+0.62)} \\ 48.72 \pm 0.11 \ \textbf{(+0.44)} \end{array}$	$\begin{array}{c} 45.00 \pm 0.08 \\ \textbf{45.90} \pm \textbf{0.07} \ (\textbf{+ 0.90}) \\ 45.59 \pm 0.09 \ (\textbf{+ 0.59}) \end{array}$	$\begin{array}{r} 45.05 \pm 0.12 \\ \textbf{45.97} \pm \textbf{0.06} \ \textbf{(+ 0.92)} \\ 45.60 \pm 0.04 \ \textbf{(+ 0.55)} \end{array}$
	MART PART-M $(s = 1)$ PART-M $(s = 10)$	$\begin{array}{r} 76.82 \pm 0.28 \\ 78.67 \pm 0.10 \ (\texttt{+} \ \texttt{1.85}) \\ \textbf{80.00} \pm \textbf{0.15} \ (\texttt{+} \ \textbf{3.18}) \end{array}$	$\begin{array}{r} 49.86 \pm 0.32 \\ \textbf{50.26} \pm \textbf{0.17} \ \textbf{(+0.40)} \\ 49.71 \pm 0.12 \ \textbf{(-0.15)} \end{array}$	$\begin{array}{r} 45.42 \pm 0.04 \\ \textbf{45.53} \pm \textbf{0.05} \ \textbf{(+ 0.11)} \\ 45.14 \pm 0.10 \ \textbf{(- 0.28)} \end{array}$	$\begin{array}{c} 45.10 \pm 0.06 \\ \textbf{45.19} \pm \textbf{0.04} \ (\textbf{+ 0.09}) \\ 44.61 \pm 0.24 \ (\textbf{- 0.49}) \end{array}$
		ŀ	ResNet-18		
SVHN	AT PART $(s = 1)$ PART $(s = 10)$	91.06 ± 0.24 $93.14 \pm 0.05 (+ 2.08)$ $93.75 \pm 0.07 (+ 2.69)$	$\begin{array}{c} 49.83 \pm 0.13 \\ \textbf{50.34} \pm \textbf{0.14} \ (\textbf{+ 0.51}) \\ 50.21 \pm 0.10 \ (\textbf{+ 0.38}) \end{array}$	47.68 ± 0.06 $48.08 \pm 0.09 (+ 0.40)$ $48.00 \pm 0.14 (+ 0.32)$	$\begin{array}{r} 45.48 \pm 0.05 \\ \textbf{45.67} \pm \textbf{0.13} \ (\textbf{+ 0.19}) \\ 45.61 \pm 0.08 \ (\textbf{+ 0.13}) \end{array}$
	TRADES PART-T $(s = 1)$ PART-T $(s = 10)$	$\begin{array}{r} 88.91 \pm 0.28 \\ 91.35 \pm 0.11 \ (+ \ 2.44) \\ \textbf{91.94} \pm \textbf{0.18} \ (+ \ \textbf{3.03}) \end{array}$	58.74 ± 0.53 $59.33 \pm 0.22 (+ 0.59)$ $59.01 \pm 0.13 (+ 0.27)$	53.29 ± 0.56 $54.04 \pm 0.16 (+ 0.75)$ $53.80 \pm 0.20 (+ 0.51)$	52.21 ± 0.47 $53.07 \pm 0.67 (+ 0.86)$ $52.61 \pm 0.24 (+ 0.40)$
	MART PART-M $(s = 1)$ PART-M $(s = 10)$	$\begin{array}{r} 89.76 \pm 0.08 \\ 91.42 \pm 0.36 \ (+ \ 1.66) \\ \textbf{93.20} \pm \textbf{0.22} \ (+ \ \textbf{3.44}) \end{array}$	58.52 ± 0.53 $58.85 \pm 0.29 (+ 0.33)$ $58.41 \pm 0.20 (- 0.11)$	52.42 ± 0.34 52.45 \pm 0.03 (+ 0.03) 52.18 \pm 0.14 (- 0.24)	$\begin{array}{c} 49.10 \pm 0.23 \\ \textbf{49.92} \pm \textbf{0.10} \ (\textbf{+ 0.82}) \\ 49.25 \pm 0.13 \ (\textbf{+ 0.15}) \end{array}$
		Wide	ResNet-34-10		
	AT PART $(s = 1)$ PART $(s = 10)$	$\begin{array}{r} 43.51 \pm 0.13 \\ 44.87 \pm 0.21 \ (+1.36) \\ \textbf{45.59} \pm \textbf{0.14} \ (+\textbf{2.08}) \end{array}$	$\begin{array}{c} 11.70 \pm 0.08 \\ \textbf{11.93} \pm \textbf{0.16} \ (\textbf{+ 0.23}) \\ 11.81 \pm 0.10 \ (\textbf{+ 0.11}) \end{array}$	$\begin{array}{c} 10.66 \pm 0.11 \\ \textbf{10.96} \pm \textbf{0.12} \ (\textbf{+ 0.30}) \\ 10.91 \pm 0.08 \ (\textbf{+ 0.25}) \end{array}$	$\begin{array}{c} 10.53 \pm 0.14 \\ \textbf{10.76} \pm \textbf{0.06} \ (\textbf{+ 0.23}) \\ 10.68 \pm 0.10 \ (\textbf{+ 0.15}) \end{array}$
TinyImagenet-200	TRADES PART-T $(s = 1)$ PART-T $(s = 10)$	$\begin{array}{c} 43.05 \pm 0.15 \\ 44.31 \pm 0.12 \ (\texttt{+} \ \texttt{1.26}) \\ \textbf{45.16} \pm \textbf{0.10} \ (\texttt{+} \ \textbf{2.11}) \end{array}$	13.86 ± 0.10 $14.08 \pm 0.22 (+ 0.22)$ $13.98 \pm 0.15 (+ 0.12)$	$\begin{array}{c} 12.62 \pm 0.16 \\ \textbf{13.01} \pm \textbf{0.09} \ \textbf{(+ 0.39)} \\ 12.88 \pm 0.12 \ \textbf{(+ 0.26)} \end{array}$	12.55 ± 0.09 12.84 ± 0.14 (+ 0.29) 12.72 ± 0.08 (+ 0.17)
	MART PART-M $(s = 1)$ PART-M $(s = 10)$	42.68 ± 0.22 $43.75 \pm 0.24 (+ 1.07)$ $45.02 \pm 0.16 (+ 2.34)$	14.77 ± 0.18 $14.93 \pm 0.15 (+ 0.16)$ $14.65 \pm 0.14 (- 0.12)$	13.58 ± 0.13 $13.76 \pm 0.06 (+ 0.18)$ $13.41 \pm 0.11 (- 0.17)$	13.42 ± 0.16 $13.68 \pm 0.13 (+ 0.24)$ $13.37 \pm 0.15 (- 0.05)$

Dataset	Method	Natural	PGD-20	MMA	AA
		ŀ	ResNet-18		
	$\begin{array}{l} \text{AT} \\ \text{PART} (s=1) \end{array}$	82.58 ± 0.14 $83.42 \pm 0.26 (+ 0.84)$	$\frac{43.69 \pm 0.28}{43.65 \pm 0.16 (-0.04)}$	41.80 ± 0.10 $41.98 \pm 0.03 (+ 0.18)$	$41.63 \pm 0.22 \\ 41.74 \pm 0.04 \ (+ \ 0.11)$
	PART $(s = 1)$ PART $(s = 10)$	$83.77 \pm 0.15 (+ 1.19)$	$43.36 \pm 0.21 (-0.33)$	$41.83 \pm 0.03 (+ 0.13)$ $41.83 \pm 0.07 (+ 0.03)$	$41.41 \pm 0.14 (-0.22)$
Dataset CIFAR-10	TRADES	78.16 ± 0.15	48.28 ± 0.05	45.00 ± 0.08	45.05 ± 0.12
	PART-T $(s = 1)$	$79.36 \pm 0.31 (+ 1.20)$	$48.90 \pm 0.14 (+0.62)$	$45.90 \pm 0.07 (+ 0.90)$	$45.97 \pm 0.06 (+ 0.92)$
	PART-T $(s = 10)$	$80.13 \pm 0.16 (+1.97)$	$48.72 \pm 0.11 (+ 0.44)$	$45.59 \pm 0.09 (+0.59)$	$45.60 \pm 0.04 \ (+ \ 0.55)$
	MART	76.82 ± 0.28	49.86 ± 0.32	45.42 ± 0.04	45.10 ± 0.06
	PART-M $(s = 1)$	$78.67 \pm 0.10 (+1.85)$	$50.26 \pm 0.17 (+ 0.40)$	$45.53 \pm 0.05 \ (+ \ 0.11)$	$45.19 \pm 0.04 \ (+ \ 0.09)$
	PART-M $(s = 10)$	$80.00 \pm 0.15 (+3.18)$	49.71 ± 0.12 (- 0.15)	45.14 ± 0.10 (- 0.28)	$44.61 \pm 0.24 (-0.49)$
		F	ResNet-18		
	AT	91.06 ± 0.24	49.83 ± 0.13	47.68 ± 0.06	45.48 ± 0.05
	PART $(s = 1)$	93.14 ± 0.05 (+ 2.08)	$50.34 \pm 0.14 (+ 0.51)$	$48.08 \pm 0.09 \ (+ \ 0.40)$	45.67 ± 0.13 (+ 0.19)
	PART $(s = 10)$	$93.75 \pm 0.07 \ (+ 2.69)$	$50.21 \pm 0.10 (+ 0.38)$	$48.00 \pm 0.14 \ (+ \ 0.32)$	$45.61 \pm 0.08 (+ 0.13)$
SVHN	TRADES	88.91 ± 0.28	58.74 ± 0.53	53.29 ± 0.56	52.21 ± 0.47
SVHN	PART-T $(s = 1)$	91.35 ± 0.11 (+ 2.44)	59.33 ± 0.22 (+ 0.59)	$54.04 \pm 0.16 \ (+ \ 0.75)$	$53.07 \pm 0.67 (+ 0.86)$
	PART-T ($s = 10$)	$91.94 \pm 0.18 (+ 3.03)$	$59.01 \pm 0.13 (+ 0.27)$	$53.80 \pm 0.20 (+ 0.51)$	$52.61 \pm 0.24 (+ 0.40)$
	MART	89.76 ± 0.08	58.52 ± 0.53	52.42 ± 0.34	49.10 ± 0.23
	PART-M $(s = 1)$	91.42 ± 0.36 (+ 1.66)	58.85 ± 0.29 (+ 0.33)	52.45 ± 0.03 (+ 0.03)	$49.92 \pm 0.10 \ (+ \ 0.82)$
	PART-M ($s = 10$)	$93.20 \pm 0.22 \ (+ \ 3.44)$	$58.41 \pm 0.20 (-0.11)$	52.18 ± 0.14 (- 0.24)	49.25 ± 0.13 (+ 0.15)
		Wide	ResNet-34-10		
	AT	43.51 ± 0.13	11.70 ± 0.08	10.66 ± 0.11	10.53 ± 0.14
	PART $(s = 1)$	$44.87 \pm 0.21 (+1.36)$	$11.93 \pm 0.16 (+ 0.23)$	$10.96 \pm 0.12 (+ 0.30)$	$10.76 \pm 0.06 (+ 0.23)$
	PART $(s = 10)$	$45.59 \pm 0.14 \ (+ \ 2.08)$	$11.81 \pm 0.10 (+ 0.11)$	$10.91 \pm 0.08 \ (+ \ 0.25)$	$10.68 \pm 0.10 \ (+ \ 0.15)$
TinyImagenet 200	TRADES	43.05 ± 0.15	13.86 ± 0.10	12.62 ± 0.16	12.55 ± 0.09
TinyImagenet-200	PART-T $(s = 1)$	$44.31 \pm 0.12 (+ 1.26)$	$14.08 \pm 0.22 \ (+ \ 0.22)$	$13.01 \pm 0.09 \ (+ \ 0.39)$	$12.84 \pm 0.14 (+ 0.29)$
	PART-T ($s = 10$)	$45.16 \pm 0.10 \ (+ \ 2.11)$	$13.98 \pm 0.15 \ (+ \ 0.12)$	$12.88 \pm 0.12 (+ 0.26)$	$12.72 \pm 0.08 (+ 0.17)$
	MART	42.68 ± 0.22	14.77 ± 0.18	13.58 ± 0.13	13.42 ± 0.16
	PART-M $(s = 1)$	43.75 ± 0.24 (+ 1.07)	$14.93 \pm 0.15 \ (+ \ 0.16)$	$13.76 \pm 0.06 \ (+ \ 0.18)$	$13.68 \pm 0.13 \ (+ \ 0.24)$
	PART-M ($s = 10$)	$45.02 \pm 0.16 \ (+ \ 2.34)$	14.65 ± 0.14 (- 0.12)	13.41 ± 0.11 (- 0.17)	13.37 ± 0.15 (- 0.05)

Dataset	Method	Natural	PGD-20	MMA	AA
		ŀ	ResNet-18		
	AT	82.58 ± 0.14	$\textbf{43.69} \pm \textbf{0.28}$	41.80 ± 0.10	41.63 ± 0.22
	PART $(s = 1)$	$83.42 \pm 0.26 (+0.84)$	43.65 ± 0.16 (- 0.04)	$41.98 \pm 0.03 \ (+ \ 0.18)$	$41.74 \pm 0.04 \ (+ \ 0.11)$
CIFAR-10	PART $(s = 10)$	$83.77 \pm 0.15 (+1.19)$	43.36 ± 0.21 (- 0.33)	$41.83 \pm 0.07 (+0.03)$	41.41 ± 0.14 (- 0.22)
	TRADES	78.16 ± 0.15	48.28 ± 0.05	45.00 ± 0.08	45.05 ± 0.12
	PART-T $(s = 1)$	79.36 ± 0.31 (+ 1.20)	$48.90 \pm 0.14 (+0.62)$	45.90 ± 0.07 (+ 0.90)	45.97 ± 0.06 (+ 0.92)
	PART-T $(s = 10)$	$80.13 \pm 0.16 \ (\texttt{+} \ \textbf{1.97})$	48.72 ± 0.11 (+ 0.44)	$45.59 \pm 0.09 \ (+ \ 0.59)$	$45.60 \pm 0.04 \ (+0.55)$
	MART	76.82 ± 0.28	49.86 ± 0.32	45.42 ± 0.04	45.10 ± 0.06
	PART-M $(s = 1)$	78.67 ± 0.10 (+ 1.85)	50.26 ± 0.17 (+ 0.40)	45.53 ± 0.05 (+ 0.11)	$45.19 \pm 0.04 (+ 0.09)$
	PART-M ($s = 10$)	$80.00 \pm 0.15 \ (\textbf{+ 3.18})$	49.71 ± 0.12 (- 0.15)	45.14 ± 0.10 (- 0.28)	44.61 ± 0.24 (- 0.49)
		ŀ	ResNet-18		
	AT	91.06 ± 0.24	49.83 ± 0.13	47.68 ± 0.06	45.48 ± 0.05
	PART $(s = 1)$	$93.14 \pm 0.05 (+ 2.08)$	$50.34 \pm 0.14 (+0.51)$	$48.08 \pm 0.09 (+0.40)$	$45.67 \pm 0.13 (+ 0.19)$
	PART $(s = 10)$	$93.75 \pm 0.07 \ (+ \ 2.69)$	$50.21 \pm 0.10 (+ 0.38)$	$48.00 \pm 0.14 (+ 0.32)$	$45.61 \pm 0.08 \ (+ \ 0.13)$
SVHN	TRADES	88.91 ± 0.28	58.74 ± 0.53	53.29 ± 0.56	52.21 ± 0.47
SVHN	PART-T $(s = 1)$	91.35 ± 0.11 (+ 2.44)	59.33 ± 0.22 (+ 0.59)	54.04 \pm 0.16 (+ 0.75)	$53.07 \pm 0.67 (+ 0.86)$
	PART-T $(s = 10)$	$91.94 \pm 0.18 (+ 3.03)$	$59.01 \pm 0.13 (+ 0.27)$	$53.80 \pm 0.20 (+0.51)$	$52.61 \pm 0.24 (+ 0.40)$
	MART	89.76 ± 0.08	58.52 ± 0.53	52.42 ± 0.34	49.10 ± 0.23
	PART-M $(s = 1)$	91.42 ± 0.36 (+ 1.66)	58.85 ± 0.29 (+ 0.33)	$52.45 \pm 0.03 \ (+ \ 0.03)$	$49.92 \pm 0.10 \ (+ \ 0.82)$
	PART-M ($s = 10$)	$93.20 \pm 0.22 (+ 3.44)$	$58.41 \pm 0.20 (-0.11)$	52.18 ± 0.14 (- 0.24)	$49.25 \pm 0.13 \ (+ \ 0.15)$
		Wide	ResNet-34-10		
	AT	43.51 ± 0.13	11.70 ± 0.08	10.66 ± 0.11	10.53 ± 0.14
TinyImagenet-200	PART $(s = 1)$	$44.87 \pm 0.21 (+1.36)$	$11.93 \pm 0.16 (+ 0.23)$	$10.96 \pm 0.12 (+ 0.30)$	$10.76 \pm 0.06 (+ 0.23)$
	PART $(s = 10)$	$45.59 \pm 0.14 \ (+ \ 2.08)$	$11.81 \pm 0.10 (+ 0.11)$	$10.91 \pm 0.08 \ (+ \ 0.25)$	$10.68 \pm 0.10 \ (+ \ 0.15)$
	TRADES	43.05 ± 0.15	13.86 ± 0.10	12.62 ± 0.16	12.55 ± 0.09
	PART-T $(s = 1)$	$44.31 \pm 0.12 (+ 1.26)$	$14.08 \pm 0.22 \ (+ \ 0.22)$	$13.01 \pm 0.09 \ (+ \ 0.39)$	$12.84 \pm 0.14 (+ 0.29)$
	PART-T $(s = 10)$	$45.16 \pm 0.10 \ (+ \ 2.11)$	$13.98 \pm 0.15 (+ 0.12)$	12.88 ± 0.12 (+ 0.26)	$12.72 \pm 0.08 (+ 0.17)$
	MART	42.68 ± 0.22	14.77 ± 0.18	13.58 ± 0.13	13.42 ± 0.16
	PART-M $(s = 1)$	$43.75 \pm 0.24 (+1.07)$	$14.93 \pm 0.15 \ (+ \ 0.16)$	$13.76 \pm 0.06 \ (+ \ 0.18)$	$13.68 \pm 0.13 \ (+ \ 0.24)$
	PART-M $(s = 10)$	$45.02 \pm 0.16 (+ 2.34)$	14.65 ± 0.14 (- 0.12)	13.41 ± 0.11 (- 0.17)	$13.37 \pm 0.15 (-0.05)$







Results

	TRADES	MART		AT	TRADES	MART	
	PART-T	PART-M		PART	PART-T	PART-M	
ch n	umber $= 30$			(b) Epoch r	number = 40		
	TRADES	MART		AT	TRADES	MART	
			- 12				
	PART-T	PART-M		PART	PART-T	PART-M	

(c) Epoch number = 50

(d) Epoch number = 60