Improving Accuracy-robustness Trade-off via Pixel Reweighted Adversarial Training

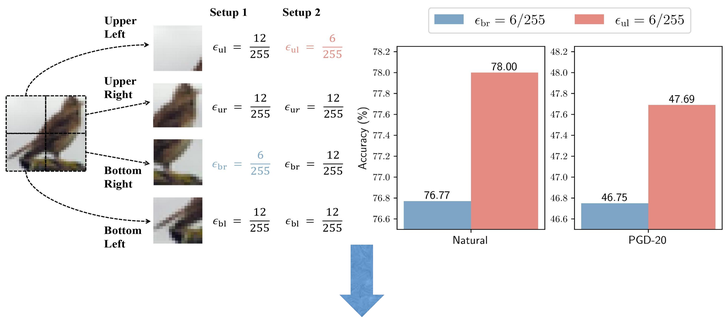
Jiacheng Zhang, Feng Liu*, Dawei Zhou, Jingfeng Zhang, Tongliang Liu*
jiachengzhang.ml@gmail.com

THE UNIVERSITY OF MELBOURNE · THE UNIVERSITY OF SYDNEY · THE UNIVERSITY OF AUCKLAND Te Whare Wānanga o Tāmaki Makaurau NEW ZEALAND
TMLR TRUSTWORTHY MACHINE LEARNING AND REASONING · ICML International Conference On Machine Learning

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.

Motivation



- 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.

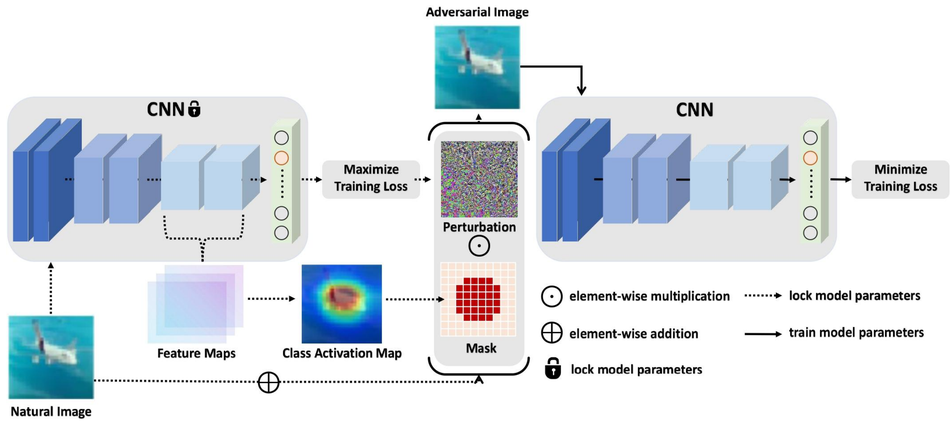
- Pixels contribute more towards higher robustness and accuracy should have more weight.

Paper **Code** **Contact**

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)

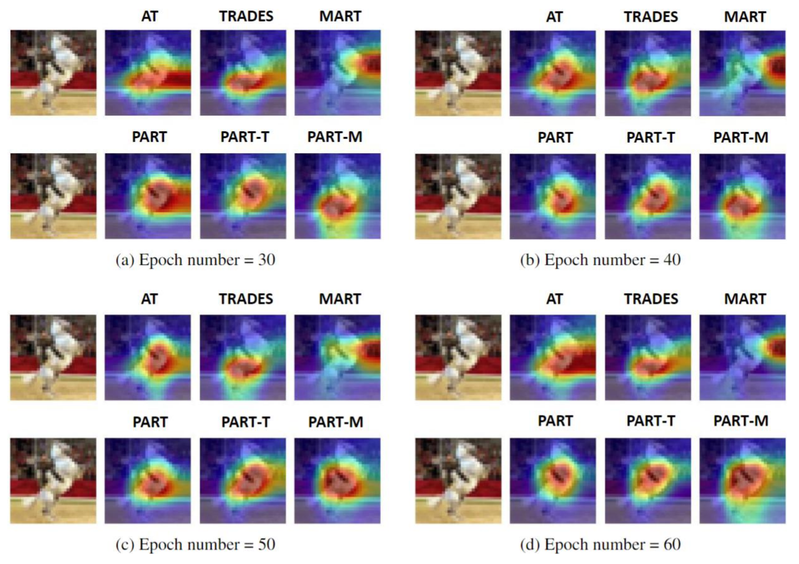


- PART leverages the power of *Class Activation Mapping* (CAM) methods to identify important pixel regions.
- PART partially reduces ϵ for less influential pixels, guiding the model to focus more on key regions that affect its outputs.
- The innovation in the generation process of AEs allows PART to 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?

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

Results



(a) Epoch number = 30 (b) Epoch number = 40

(c) Epoch number = 50 (d) Epoch number = 60

Dataset	Method	Natural	PGD-20	MMA	AA
		ResNet-18			
CIFAR-10	AT	82.58 ± 0.14	**43.69 ± 0.28**	41.80 ± 0.10	41.63 ± 0.22
	PART ($s = 1$)	83.42 ± 0.26 (+ 0.84)	43.65 ± 0.16 (- 0.04)	**41.98 ± 0.03 (+ 0.18)**	**41.74 ± 0.04 (+ 0.11)**
	PART ($s = 10$)	**83.77 ± 0.15 (+ 1.19)**	43.36 ± 0.21 (- 0.33)	41.83 ± 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 ± 0.14 (+ 0.62)**	**45.90 ± 0.07 (+ 0.90)**	**45.97 ± 0.06 (+ 0.92)**
	PART-T ($s = 10$)	**80.13 ± 0.16 (+ 1.97)**	48.72 ± 0.11 (+ 0.44)	45.59 ± 0.09 (+ 0.59)	45.60 ± 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 ± 0.04 (+ 0.09)**
	PART-M ($s = 10$)	**80.00 ± 0.15 (+ 3.18)**	49.71 ± 0.12 (- 0.15)	45.14 ± 0.10 (- 0.28)	44.61 ± 0.24 (- 0.49)
		ResNet-18			
SVHN	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 ± 0.14 (+ 0.51)**	**48.08 ± 0.09 (+ 0.40)**	**45.67 ± 0.13 (+ 0.19)**
	PART ($s = 10$)	**93.75 ± 0.07 (+ 2.69)**	50.21 ± 0.10 (+ 0.38)	48.00 ± 0.14 (+ 0.32)	45.61 ± 0.08 (+ 0.13)
	TRADES	88.91 ± 0.28	58.74 ± 0.53	53.29 ± 0.56	52.21 ± 0.47
	PART-T ($s = 1$)	91.35 ± 0.11 (+ 2.44)	**59.33 ± 0.22 (+ 0.59)**	**54.04 ± 0.16 (+ 0.75)**	**53.07 ± 0.67 (+ 0.86)**
	PART-T ($s = 10$)	**91.94 ± 0.18 (+ 3.03)**	59.01 ± 0.13 (+ 0.27)	53.80 ± 0.20 (+ 0.51)	52.61 ± 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 ± 0.10 (+ 0.82)**
	PART-M ($s = 10$)	**93.20 ± 0.22 (+ 3.44)**	58.41 ± 0.20 (- 0.11)	52.18 ± 0.14 (- 0.24)	49.25 ± 0.13 (+ 0.15)
		WideResNet-34-10			
TinyImagenet-200	AT	43.51 ± 0.13	11.70 ± 0.08	10.66 ± 0.11	10.53 ± 0.14
	PART ($s = 1$)	44.87 ± 0.21 (+ 1.36)	**11.93 ± 0.16 (+ 0.23)**	**10.96 ± 0.12 (+ 0.30)**	**10.76 ± 0.06 (+ 0.23)**
	PART ($s = 10$)	**45.59 ± 0.14 (+ 2.08)**	11.81 ± 0.10 (+ 0.11)	10.91 ± 0.08 (+ 0.25)	10.68 ± 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 ± 0.12 (+ 1.26)	**14.08 ± 0.22 (+ 0.22)**	**13.01 ± 0.09 (+ 0.39)**	**12.84 ± 0.14 (+ 0.29)**
	PART-T ($s = 10$)	**45.16 ± 0.10 (+ 2.11)**	13.98 ± 0.15 (+ 0.12)	12.88 ± 0.12 (+ 0.26)	12.72 ± 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 ± 0.15 (+ 0.16)**	**13.76 ± 0.06 (+ 0.18)**	**13.68 ± 0.13 (+ 0.24)**
	PART-M ($s = 10$)	**45.02 ± 0.16 (+ 2.34)**	14.65 ± 0.14 (- 0.12)	13.41 ± 0.11 (- 0.17)	13.37 ± 0.15 (- 0.05)