IEEE ICASSP 2022 Paper No. 1484

DEFENDING AGAINST UNIVERSAL ATTACK VIA CURVATURE-AWARE CATEGORYADVERSARIALTRAINING

Peilun Du, Xiaolong Zheng, Liang Liu, Huadong Ma

Beijing Key Lab of Intelligent Telecomm. Software and Multimedia, Beijing University of Posts and Telecomm.

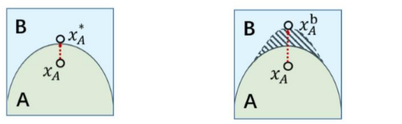
Introduction

Background

- ✓ Adversarial training methods defend against universal adversarial perturbation (UAP) by injecting corresponding adversarial samples during training.
- ✓ Training with UAP inevitably includes excessive perturbations related to other categories.
- ✓ High training cost hinders the application of adversarial training.

Observation

- ✓ Training with UAP will cause more erroneous predictions with larger local positive curvature.
- ✓ The geometric argument demonstrate the excessive perturbations.



(a) Desired sample x_A^* closes to the decision boundary. (b) Excessive perturbations x_A^b cause larger positive curvature.

Figure 1: Geometric illustration of curvature during adversarial training.

	FGSM	PGD	Deepfool	F-UAP	Ours
RC	10.56	11.79	14.50	12.02	**9.79**
BD	1.68	1.48	**1.09**	-	1.38

Table 1: Geometric arguments of adversaries on CIFAR10.

Challenges & Solutions

- ✓ Excessive perturbations of training samples
  Solution: Category-oriented adversarial mask
- ✓ Local positive curvature of decision boundary
  Solution: Curvature-aware adversarial training
- ✓ High training cost
  Solution: Splitting the min-max optimization loops

Method

Framework

- ✓ To defend against universal adversarial perturbation, we propose a curvature-aware adversarial training framework.
- ✓ We generate category-oriented adversarial mask with cumulative momentum.
- ✓ We split the min-max optimization loops of adversarial training into two parallel processes to reduce the training cost.

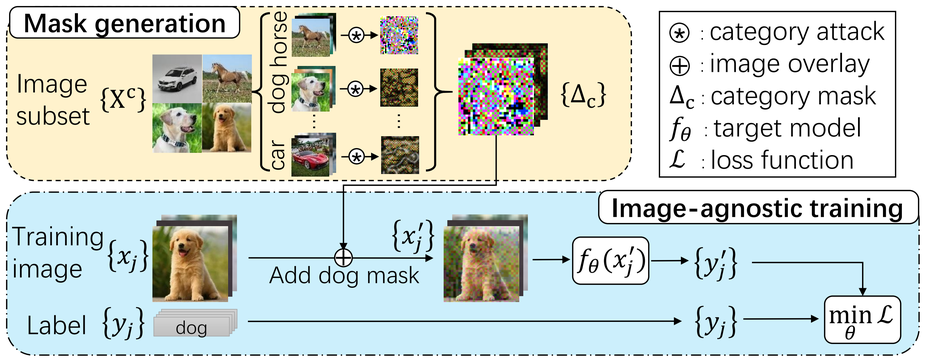


Figure 2: The overview of curvature-aware adversarial training.

Image-agnostic Adversarial Training

- ✓ The training process include two parallel stages.
- ✓ We can generate CoAM for the next epoch in advance.

$$\min_\theta \sum \max_{\delta<\epsilon} \mathcal{L}(f_\theta(x+\delta^*), y)$$

$$\begin{cases} \max\limits_{\Delta_c<\epsilon} \sum\limits_{x\in\{X^c\}} \mathcal{L}(f_\theta(x+\delta), y) & stage_1 \\ min_\theta \sum \mathcal{L}(f_\theta(x+\Delta_c), y) & stage_2, \end{cases}$$

Experiments

	Clean	Madry	SAT	UAT	Ours
BD	1.04	1.19	1.07	0.25	**1.42**
Cost (s)	-	-	-	2819	**316**

Table 2: Comparison of robust distance and training cost.

		Clean	C-UAP	F-UAP	CD-UAP
Models trained with	Clean	96.4	11.7	13.7	10.2
	Madry	88.7	83.2	85.6	81.9
	SAT	93.2	86.5	88.7	84.4
	UAT	93.5	93.3	91.8	89.6
	Ours	**94.4**	**94.0**	**93.3**	**92.6**

Table 3: Accuracy (%) comparison on CIFAR-10.

Conclusion

- ✓ We analyze the geometric arguments of adversaries in existing adversarial training on CIFAR10.
- ✓ We propose a curvature-aware adversarial training framework which trained with CoAM.
- ✓ We utilize a parallel training method by splitting the min-max optimization loops of adversarial training to reduce the training cost.