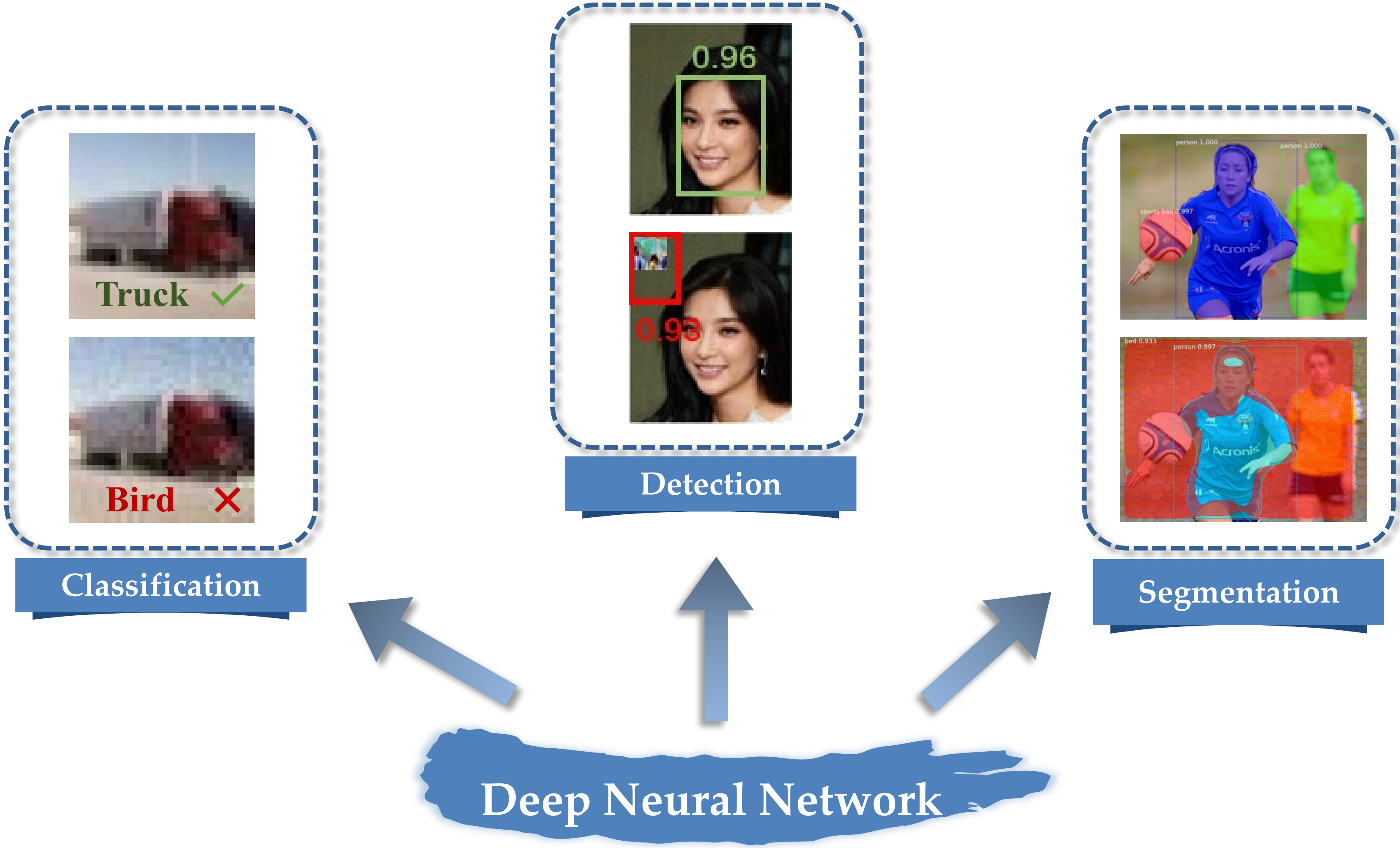
Improving Adversarial Robustness via Mutual Information Estimation

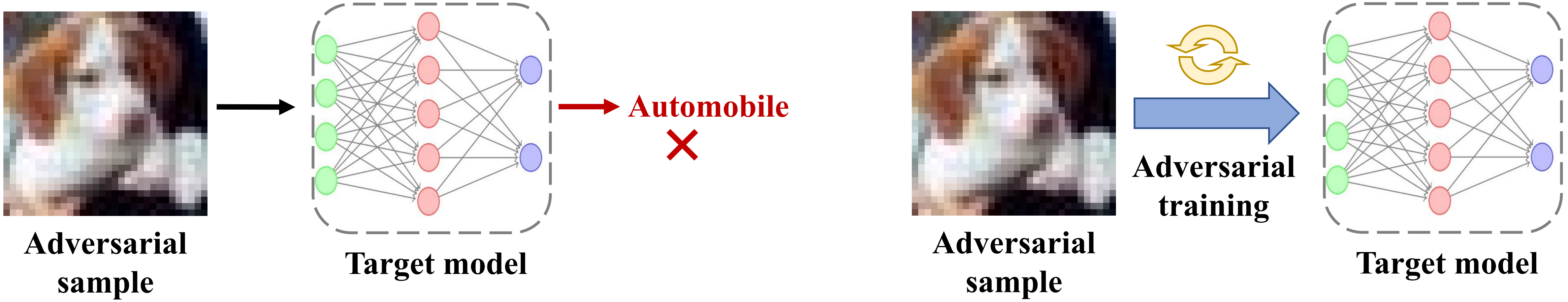
Dawei Zhou[1], Nannan Wang[1 †], Xinbo Gao[2], Bo Han[3], Xiaoyu Wang[4], Yibing Zhan[5], Tongliang Liu[6]

[1]Xidian University, [2]Chongqing University of Posts and Telecommunications, [3]Hong Kong Baptist University,
[4]The Chinese University of Hong Kong (Shenzhen), [5]JD Explore Academy, [6]University of Sydney

Introduction

Deep neural networks are vulnerable to adversarial noise

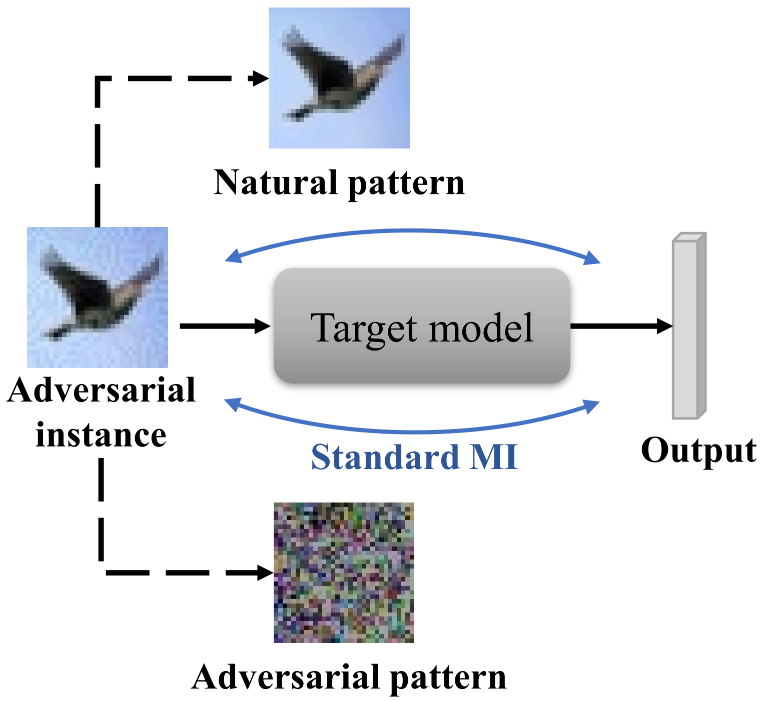




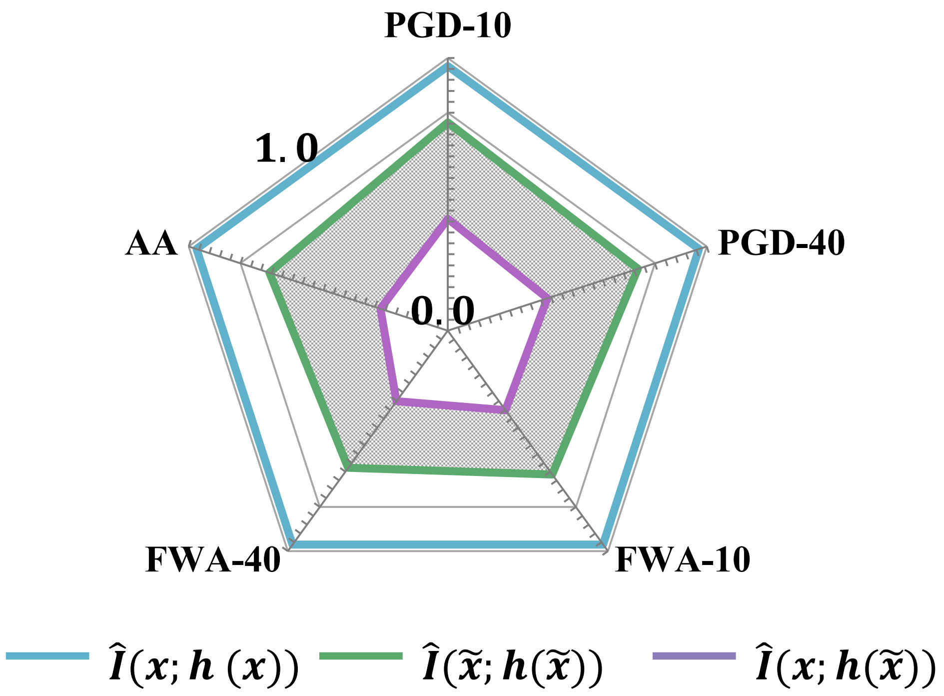
- The dependence between the input adversarial sample and the corresponding output and has not been well studied yet.
- Explicitly measuring this dependence could help train the target model to make predictions that are more closely relevant to the ground-truth objectives.

Motivation

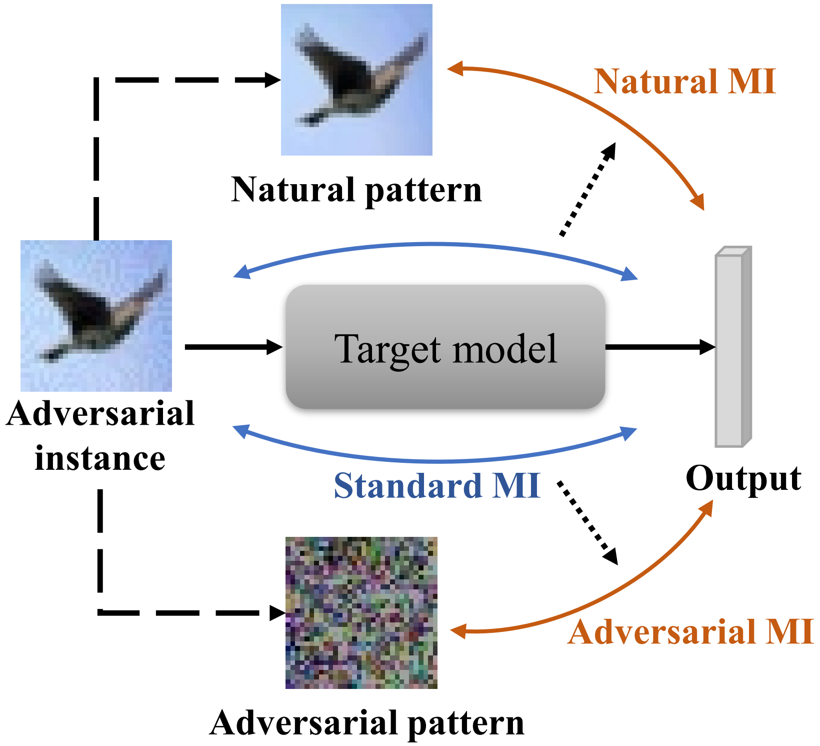
- We exploit mutual information (MI) to explicitly measure the dependence of the output on the adversarial sample.
- Adversarial samples have two patterns: the natural pattern and the adversarial pattern.
- The standard MI (i.e., MI between the adversarial sample and its corresponding output) cannot respectively consider the dependence of the output on the different patterns.



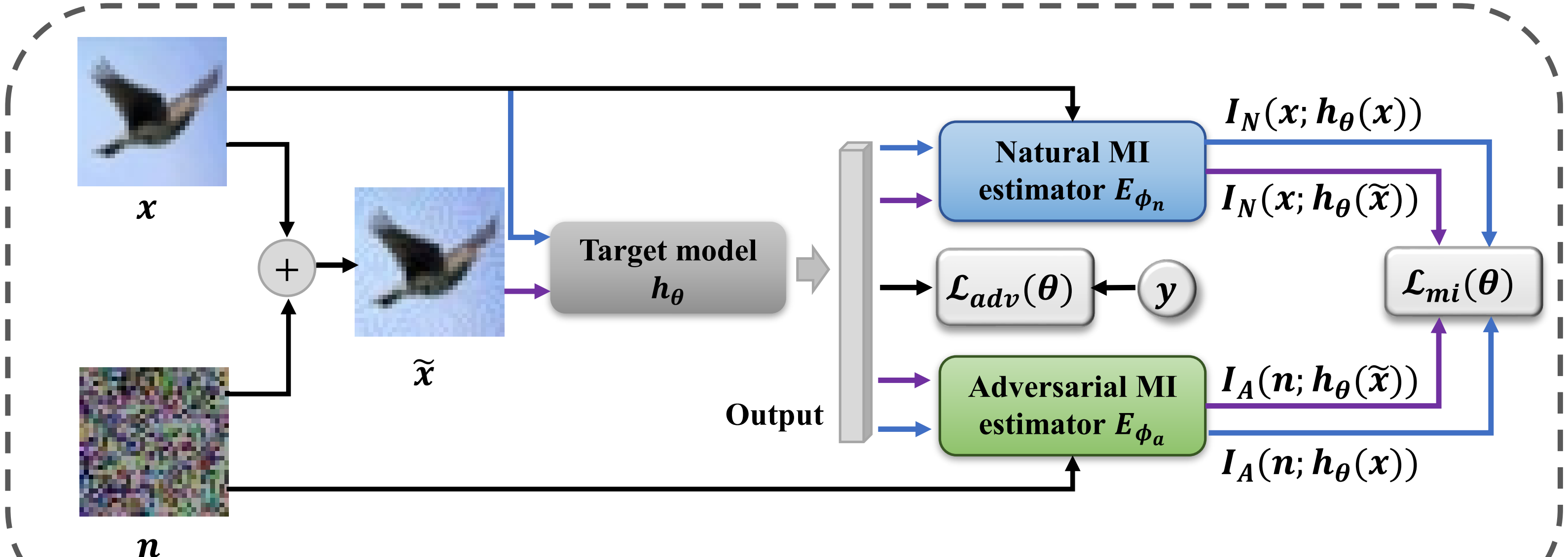
Motivation



- Standard MI of the adversarial sample contains the dependence of the output on the adversarial pattern.
- Maximizing the standard MI may increase the dependence of the output on the adversarial pattern and cause more disturbance to the prediction.



- Disentangle standard MI into natural MI (i.e., MI between outputs and natural patterns of inputs) and adversarial MI (i.e., MI between outputs and adversarial patterns of inputs).



Natural-adversarial mutual information-based defense (NAMID)

- Guide the target model to increase the attention to the natural pattern while reducing the attention to the adversarial pattern.
- Maximize the natural MI of the input adversarial sample and minimize its adversarial MI simultaneously.

- Defending against white-box attacks.

Table.1 Adversarial accuracy (higher is better) of defense methods against white-box attacks on CIFAR-10 and Tiny-ImageNet. The target model is ResNet-18.

Dataset	Defense	L_∞-norm					L_2-norm			
		None	PGD-40	AA	FWA-40	TI-DIM	None	PGD-40	CW	DDN
CIFAR-10	Standard	83.39	42.38	39.01	15.44	55.63	83.97	61.69	30.96	29.34
	WMIM	80.32	40.76	36.05	12.14	53.10	81.29	58.36	28.41	27.13
	NAMID	**83.41**	**44.79**	**39.26**	**15.67**	**58.23**	**84.35**	**62.38**	**34.48**	**32.41**
	TRADES	**80.70**	46.29	42.71	20.54	57.06	83.72	63.17	33.81	32.06
	NAMID_T	80.67	**47.53**	**43.39**	**21.17**	**59.13**	**84.19**	**64.75**	**35.41**	**34.27**
	MART	78.21	50.23	43.96	25.56	58.62	83.36	65.38	35.57	34.69
	NAMID_M	**78.38**	**51.69**	**44.42**	**25.64**	**61.26**	**84.07**	**66.03**	**36.19**	**35.76**
Tiny-ImageNet	Standard	48.40	17.35	11.27	10.29	27.84	49.57	26.19	12.73	11.25
	WMIM	47.43	16.50	9.87	9.25	25.19	48.16	24.10	11.35	10.16
	NAMID	**48.41**	**18.67**	**12.29**	**11.32**	**29.37**	**49.65**	**28.13**	**14.29**	**12.57**
	TRADES	**48.25**	19.17	12.36	10.67	29.64	48.83	27.16	13.28	12.34
	NAMID_T	48.21	**20.12**	**12.86**	**14.91**	**30.81**	**49.07**	**28.83**	**14.47**	**13.91**
	MART	**47.83**	20.90	15.57	12.95	30.71	48.56	27.98	14.36	13.79
	NAMID_M	47.80	**21.23**	**15.83**	**15.09**	**31.59**	**48.72**	**29.14**	**15.06**	**14.23**

WMIM: A defense method that combines adversarial training with standard MI maximization.

- Defending against black-box attacks.

Table.2 Adversarial accuracy (higher is better) of defense methods against black-box attacks on CIFAR-10. The target model is ResNet-18 and the surrogate model is adversarially trained VggNet-19.

Defense	None	PGD-40	AA	FWA-40
Standard	83.39	65.88	60.93	56.42
WMIM	80.32	62.79	57.86	53.05
NAMID	**83.41**	**69.57**	**63.72**	**59.30**

WMIM: A defense method that combines adversarial training with standard MI maximization.

Thank You