Improving Adversarial Robustness via Promoting Ensemble Diversity

Tianyu Pang, Kun Xu, Chao Du, Ning Chen and Jun Zhu

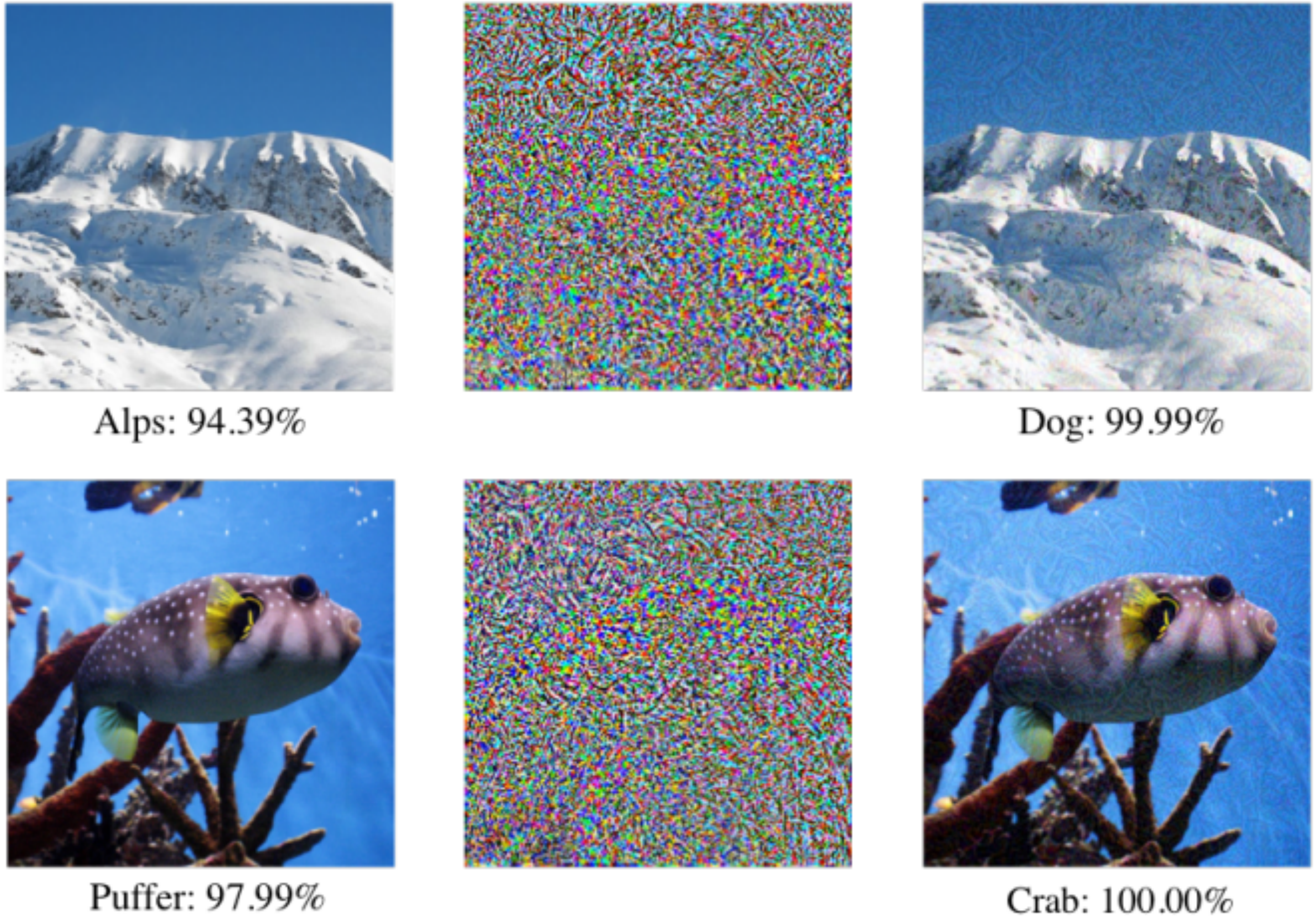
Department of Computer Science and Technology
Tsinghua University

清華大學 Tsinghua University

ICML | 2019

TSAIL

Adversarial Examples

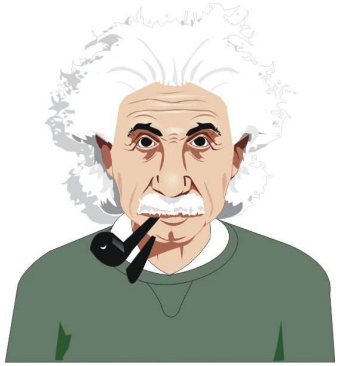


From Dong et al. (CVPR 2018)

Previous Defense Strategies

Single model defense:

e.g., adversarial training

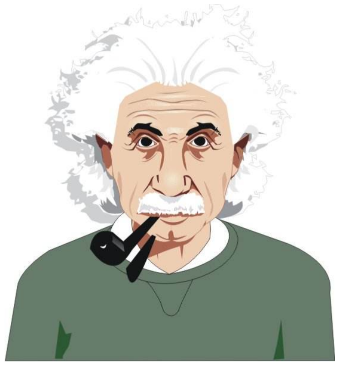


Base Model

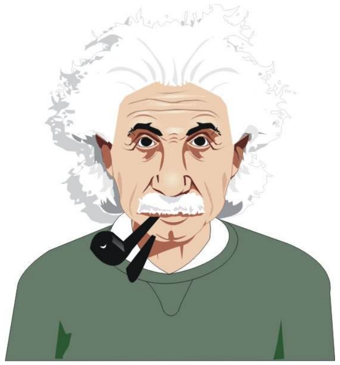
Enhanced Model

Previous Defense Strategies

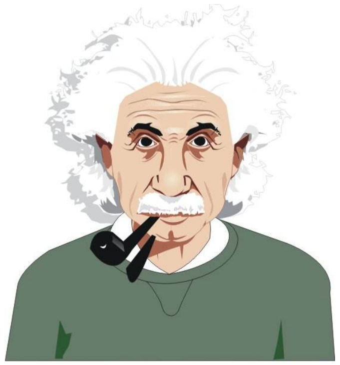
Ensemble model defense:



Member 1



Member 2

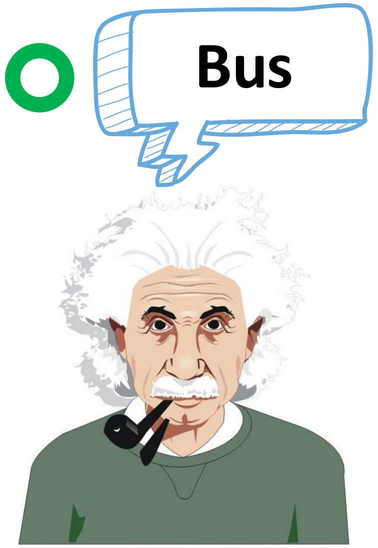


Member 3

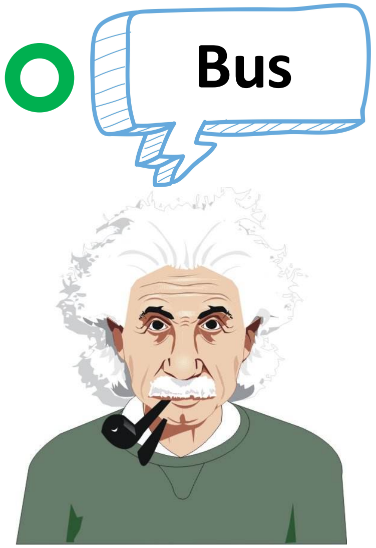
Previous Defense Strategies

Ensemble model defense:

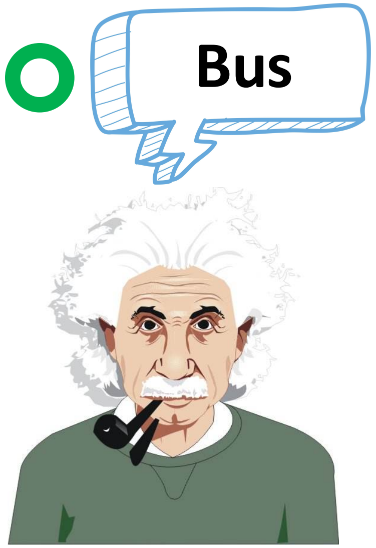
Clean input



Member 1



Member 2



Member 3

Previous Defense Strategies

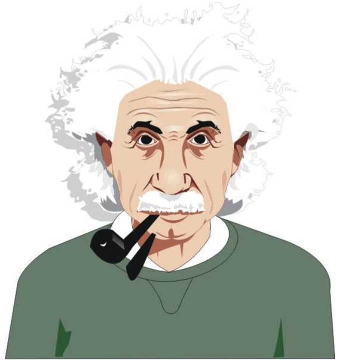
Ensemble model defense:

Member 1 Member 2 Member 3

Adversarial input

Our Strategy

Training ensembles with diversity:



Member 1

Member 2

Member 3

Our Strategy

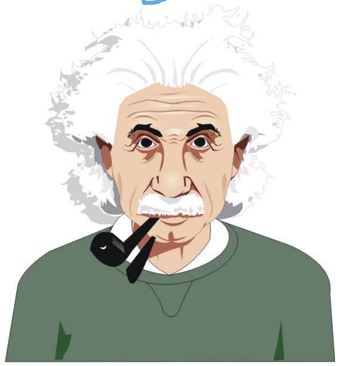
Training ensembles with diversity:

Adversarial input

Cat

Bus

Bus

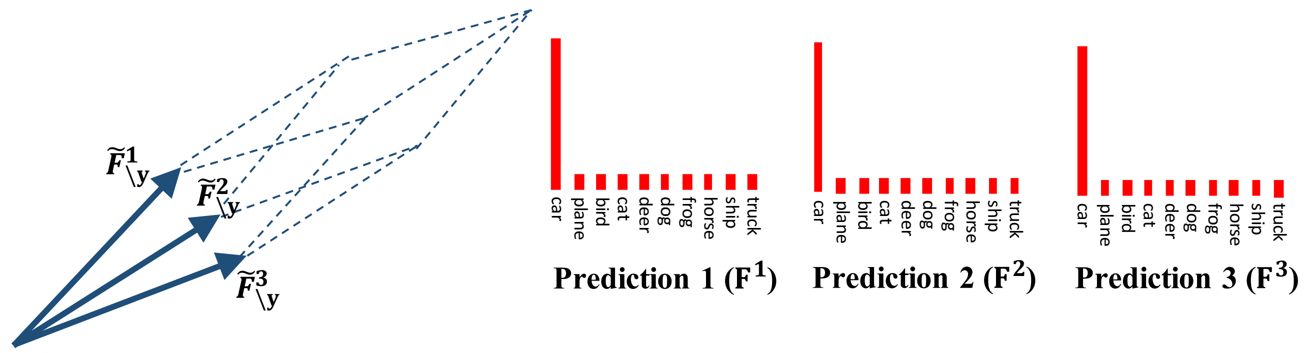


Member 1

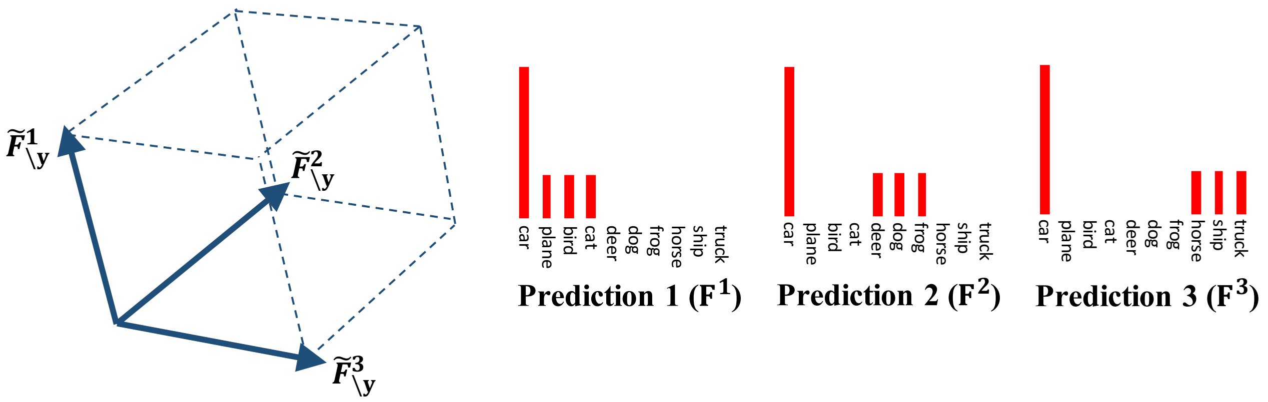
Member 2

Member 3

Adaptive Diversity Promoting



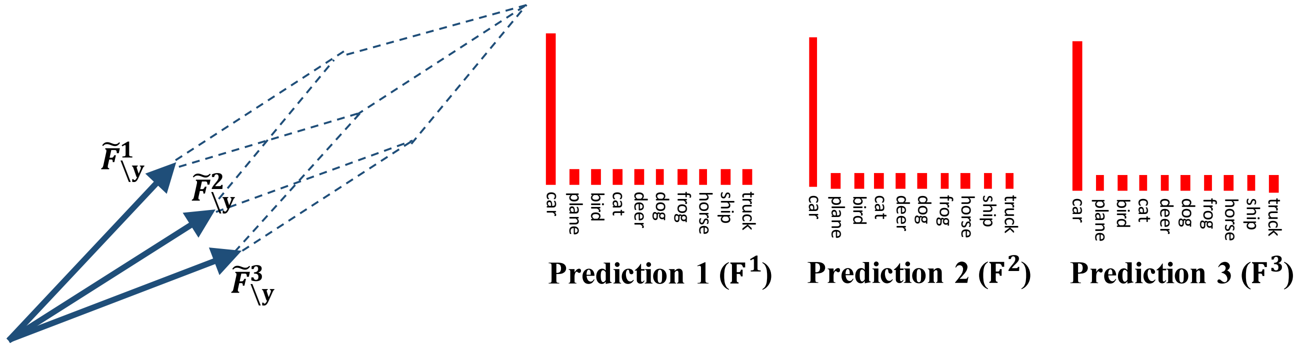
Baseline



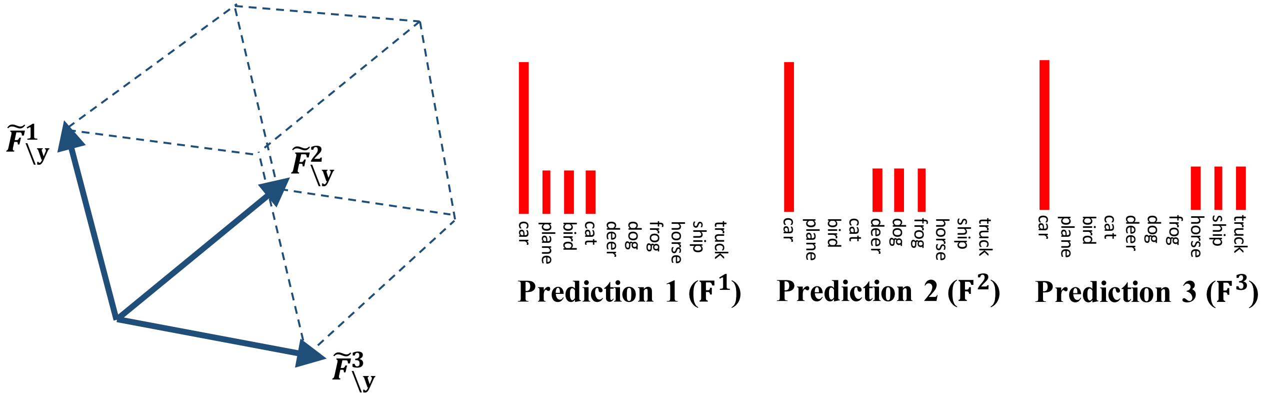
ADP

- Promoting diversity on **non-maximal predictions**

Adaptive Diversity Promoting



Baseline

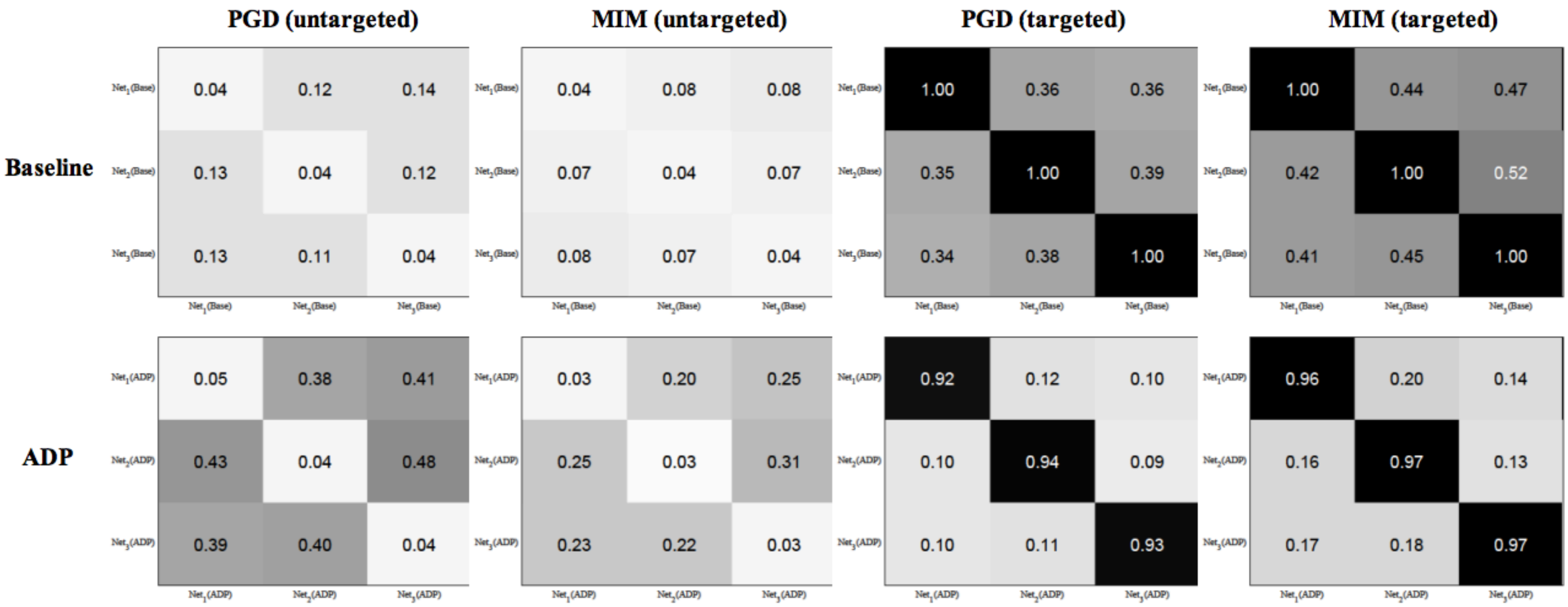


ADP

- Promoting diversity on **non-maximal predictions**

correspond to all potentially wrong labels returned for the adversarial examples

Experiments



Adversarial transferability among individual members of ensembles

Experiments

Table 2. Classification accuracy (%) on adversarial examples. Ensemble models consist of three Resnet-20. For JSMA, the perturbation $\epsilon = 0.2$ on MNIST, and $\epsilon = 0.1$ on CIFAR-10. For EAD, the factor of L_1-norm $\beta = 0.01$ on both datasets.

Attacks	MNIST				CIFAR-10			
	Para.	Baseline	$ADP_{2,0}$	$ADP_{2,0.5}$	Para.	Baseline	$ADP_{2,0}$	$ADP_{2,0.5}$
FGSM	$\epsilon = 0.1$	78.3	95.5	**96.3**	$\epsilon = 0.02$	36.5	57.4	**61.7**
	$\epsilon = 0.2$	21.5	50.6	**52.8**	$\epsilon = 0.04$	19.4	41.9	**46.2**
BIM	$\epsilon = 0.1$	52.3	86.4	**88.5**	$\epsilon = 0.01$	18.5	44.0	**46.6**
	$\epsilon = 0.15$	12.2	69.5	**73.6**	$\epsilon = 0.02$	6.1	28.2	**31.0**
PGD	$\epsilon = 0.1$	50.7	73.4	**82.8**	$\epsilon = 0.01$	23.4	43.2	**48.4**
	$\epsilon = 0.15$	6.3	36.2	**41.0**	$\epsilon = 0.02$	6.6	26.8	**30.4**
MIM	$\epsilon = 0.1$	58.3	89.7	**92.0**	$\epsilon = 0.01$	23.8	49.6	**52.1**
	$\epsilon = 0.15$	16.1	73.3	**77.5**	$\epsilon = 0.02$	7.4	32.3	**35.9**
JSMA	$\gamma = 0.3$	84.0	88.0	**95.0**	$\gamma = 0.05$	29.5	33.0	**43.5**
	$\gamma = 0.6$	74.0	85.0	**91.0**	$\gamma = 0.1$	27.5	32.0	**37.0**
C&W	$c = 0.1$	91.6	95.9	**97.3**	$c = 0.001$	71.3	76.3	**80.6**
	$c = 1.0$	30.6	75.0	**78.1**	$c = 0.01$	45.2	50.3	**54.9**
	$c = 10.0$	5.9	20.2	**23.8**	$c = 0.1$	18.8	19.2	**25.6**
EAD	$c = 5.0$	29.8	91.3	**93.4**	$c = 1.0$	17.5	64.5	**67.3**
	$c = 10.0$	7.3	87.4	**89.5**	$c = 5.0$	2.4	23.4	**29.6**

Classification accuracy (%) on adversarial examples

Experiments

Table 4. Classification accuracy (%): **AdvT$_{\text{FGSM}}$** denotes adversarial training (AdvT) on FGSM, **AdvT$_{\text{PGD}}$** denotes AdvT on PGD. $\epsilon = 0.04$ for FGSM; $\epsilon = 0.02$ for BIM, PGD and MIM.

Defense Methods	CIFAR-10			
	FGSM	BIM	PGD	MIM
AdvT$_{\text{FGSM}}$	39.3	19.9	24.2	24.5
AdvT$_{\text{FGSM}}$ + ADP$_{2,0.5}$	**56.1**	**25.7**	**26.7**	**30.6**
AdvT$_{\text{PGD}}$	43.2	27.8	32.8	32.7
AdvT$_{\text{PGD}}$ + ADP$_{2,0.5}$	**52.8**	**34.0**	**36.2**	**38.8**

Classification accuracy (%) on adversarial examples

For more technical details and results, please come

Poster:
#64

Code:
https://github.com/P2333