ICLR

Adversarial Training on Purification (AToP)

Guang Lin, Chao Li, Jianhai Zhang, Toshihisa Tanaka, Qibin Zhao*

RIKEN

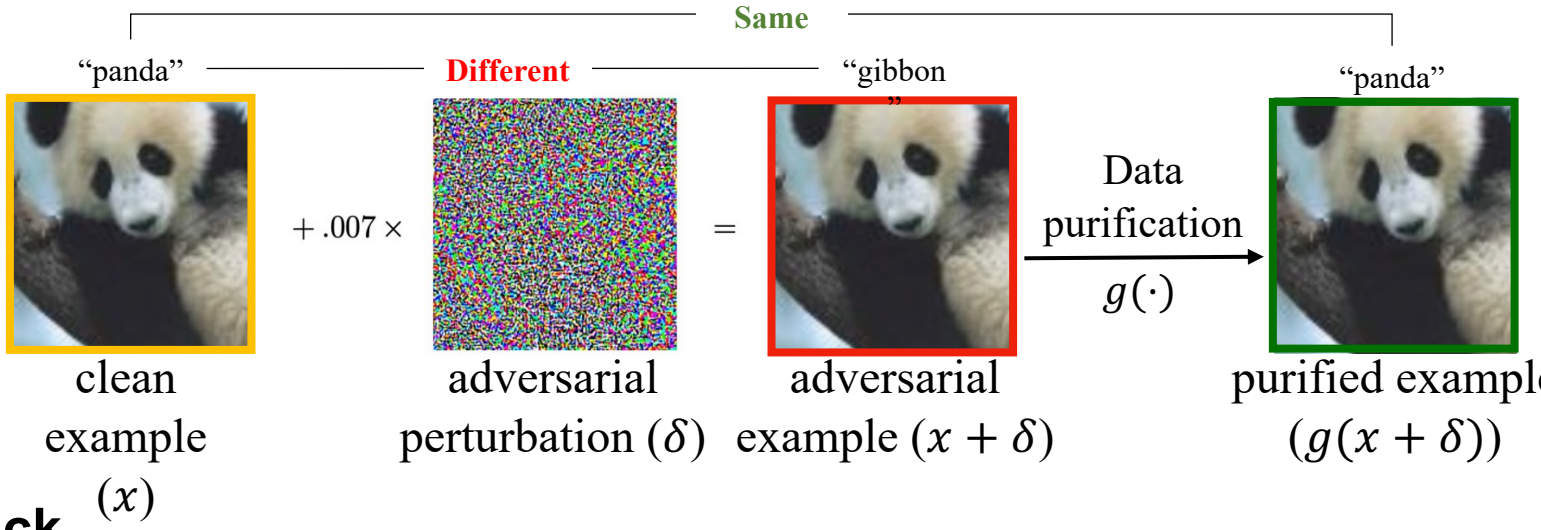
Programs for Junior Scientists

Guang Lin <guang.lin@riken.jp>, Junior Research Associate (JRA) & Ph.D.
RIKEN AIP & Tokyo University of Agriculture and Technology (TUAT)

HANGZHOU DIANZI UNIVERSITY

TAT

Attack and Defense



Adversarial attack

- An image containing a small perturbation (δ) to human but completely changes the prediction results to classifier model (f).

$$\Rightarrow \quad f(x) \neq f(x+\delta) \qquad (1)$$

Adversarial training (AT)

- Train a new robust model (f') with adversarial examples and true label y.

$$\Rightarrow \quad f(\cdot) \rightarrow f'(\cdot), \quad f'(x+\delta) = f'(x) = y \qquad (2)$$

Adversarial purification (AP) as illustrated in the diagram.

- Use pre-trained g to purify adversarial examples before classification.

$$\Rightarrow \quad f(g(x)) = f\big(g(x+\delta)\big) = y \qquad (3)$$

The figure is modified based on: Explaining and Harnessing Adversarial Examples. ICLR 2015.

Related works

Table 1: Accuracy comparison of defenses with vanilla model (negative impacts are marked in red).

Defense method	Clean examples	Known attacks	Unseen attacks	Training cost
Vanilla model	~94%	~0%	~0%	/
Expectation	=	↑↑↑	↑↑	/
AT	↓↓	↑↑↑	≈	↑
AP	↓	↑↑	↑↑	/
AToP (Ours)	≈	↑↑↑	↑↑	↑

Adversarial Training (AT)

[✓] Achieve optimal robustness on known attacks.

[✗] Vulnerable to unseen attacks.
[✗] Reduce the accuracy of clean examples.

Adversarial Purification (AP)

[✓] Keep generalization against unseen attacks.

[✗] Weaker robustness than AT on known attacks.
[✗] Slightly reduce the accuracy of clean examples.

In AP, the pre-trained generator is not good enough for classification and robustness.

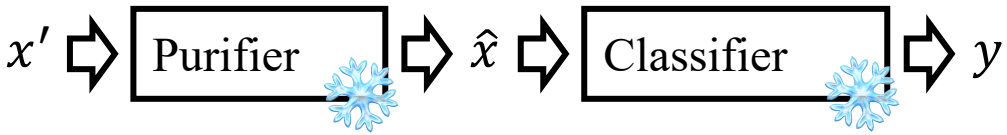
➡ Fine-tuning the purifier model with adversarial loss.

➡ Adversarial Training on Purification (AToP).

Adversarial Training on Purification (AToP)

AT: **Learning** a robust **classifier**.

AP: **Utilizing** a pre-trained generator as **purifier**.



Pre-training generator-based model:

$$L_{\theta_g} = L_g(x, \theta_g).$$

Fine-tuning generator-based model
with clean examples:

$$L_{\theta_g} = L_g(x, \theta_g) + \lambda \cdot L_{cls}(x, y, \theta_g, \theta_f),$$

with adversarial examples:

$$L_{\theta_g} = L_g(x', \theta_g) + \lambda \cdot L_{cls}(x', y, \theta_g, \theta_f).$$

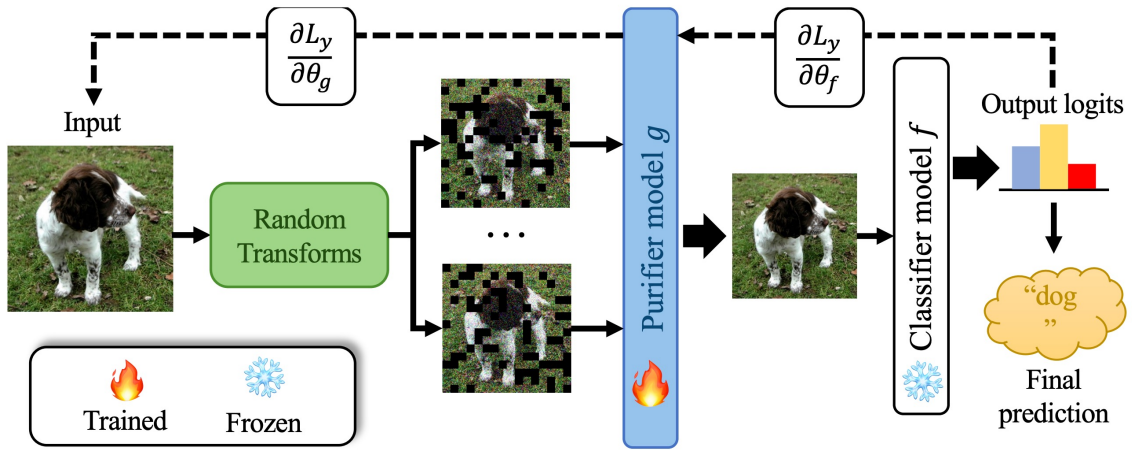
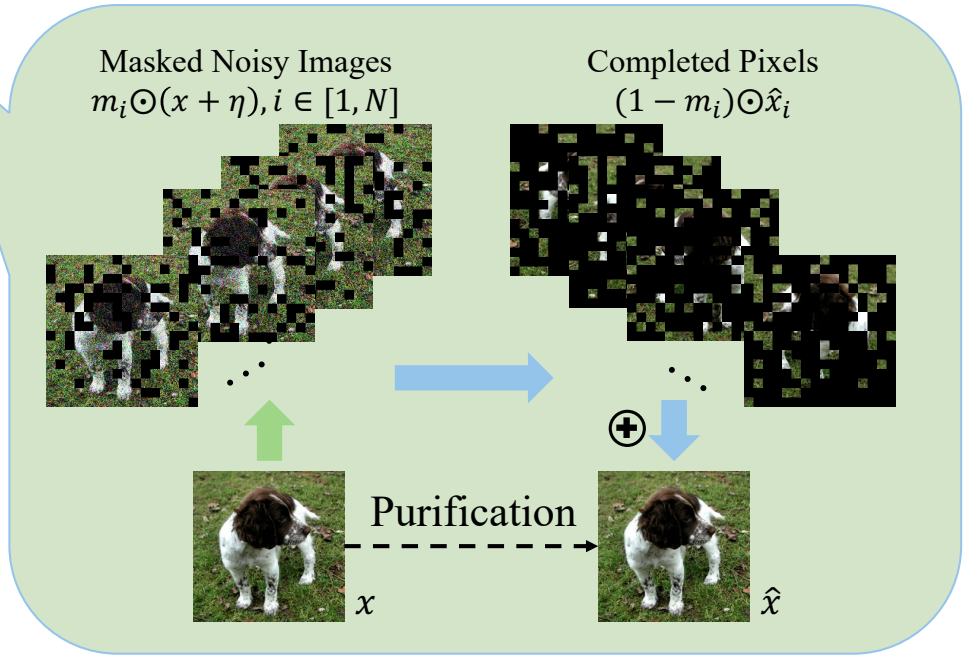


Figure 1: Illustration of adversarial training on purification (AToP): **Learning** a robust **purifier**.



Experiments

Table 6: Standard accuracy and robust accuracy against AutoAttack l_∞ ($\epsilon = 8/255$), l_2 ($\epsilon = 1$) and StAdv non-l_p ($\epsilon = 0.05$) threat models on CIFAR-10 with ResNet-50 classifier model. We utilize GAN-based model with RT_2, and all settings follow the same as Laidlaw et al. (2021).

Defense method	Standard Acc.	l_∞	l_2	StAdv
Standard Training	94.8	0.0	0.0	0.0
Adv. Training with l_∞ (Laidlaw et al., 2021)	86.8	49.0	19.2	4.8
Adv. Training with l_2 (Laidlaw et al., 2021)	85.0	39.5	47.8	7.8
Adv. Training with StAdv (Laidlaw et al., 2021)	86.2	0.1	0.2	53.9
Adv. Training with all (Laidlaw et al., 2021)	84.0	25.7	30.5	40.0
PAT-self (Laidlaw et al., 2021)	82.4	30.2	34.9	46.4
Adv. CRAIG (Dolatabadi et al., 2022)	83.2	40.0	33.9	49.6
DiffPure (Nie et al., 2022)	88.2	70.0	70.9	55.0
Ours	**89.1**	**71.2**	**73.4**	**56.4**

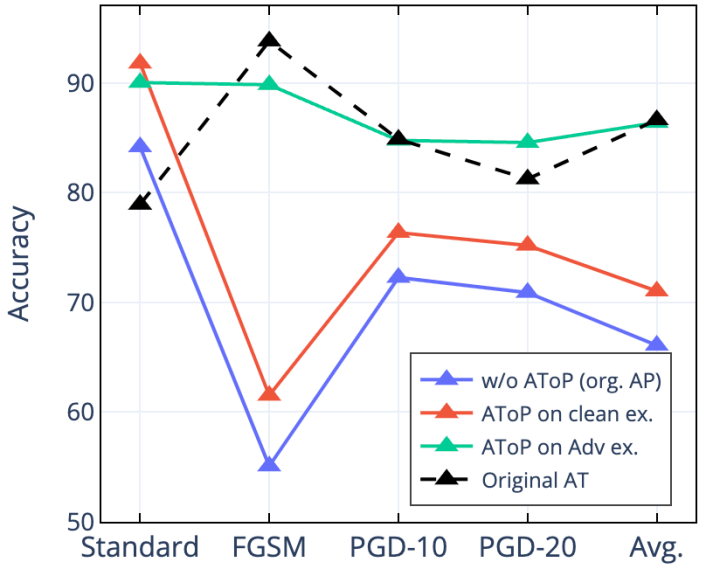


Figure 2: Standard acc. and robust acc. against l_∞ ($\epsilon = 8/255$) threat with ResNet-18.

Adversarial Training on Purification (AToP)

[✓] Achieve optimal robustness on known attacks.
[✓] Keep generalization against unseen attacks.
[✓] Achieve optimal accuracy on clean examples.

Limitations [✗] **:** AToP requires training on the purifier model, and as the complexity of the purifier model increases, so does the training cost.

Thanks for your attention!

Guang Lin <guang.lin@riken.jp>, Junior Research Associate (JRA) & Ph.D.
RIKEN AIP & Tokyo University of Agriculture and Technology (TUAT)