Adversarial Distributional Training for Robust Deep Learning

Yinpeng Dong*, Zhijie Deng*, Tianyu Pang, Hang Su, Jun Zhu

Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University, Beijing, 100084 China

{dyp17, dzj17, pty17}@mails.tsinghua.edu.cn, {suhangss, dcszj}@mail.tsinghua.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Adversarial training (AT) is among the most effective techniques to improve model robustness, which can be formulated as

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta_i \in S} L(f_\theta(x_i + \delta_i), y_i)\,,$$

where f_θ is the DNN, L is a loss function (e.g., cross-entropy loss), and $S = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ is a perturbation set.

The inner problem can be solved by projected gradient descent (Madry et al., 2018) as

$$\delta_i^{t+1} = \Pi_S\left(\delta_i^t + \alpha \cdot \mathrm{sign}\left(\nabla_x L(f_\theta(x_i + \delta_i^t), y_i)\right)\right).$$

Adversarial Distributional Training (ADT)

ADT models the adversarial perturbations around each natural example x_i by a distribution $p(\delta_i)$ as

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{p(\delta_i) \in P} \mathbb{E}_{p(\delta_i)}[L(f_\theta(x_i + \delta_i), y_i)]\,.$$

The inner maximization aims to learn an adversarial distribution, such that a point drawn from it is likely an adversarial example.
The outer minimization aims to adversarially train the model parameters by minimizing the expected loss over the worst-case adversarial distributions.

Note that

$$\max_{p(\delta_i) \in P} \mathbb{E}_{p(\delta_i)}[L(f_\theta(x_i + \delta_i), y_i)] \leq \max_{\delta_i \in S} L(f_\theta(x_i + \delta_i), y_i)$$

indicating that ADT will degenerate into AT.
Therefore, we add an entropic regularization term into the objective as

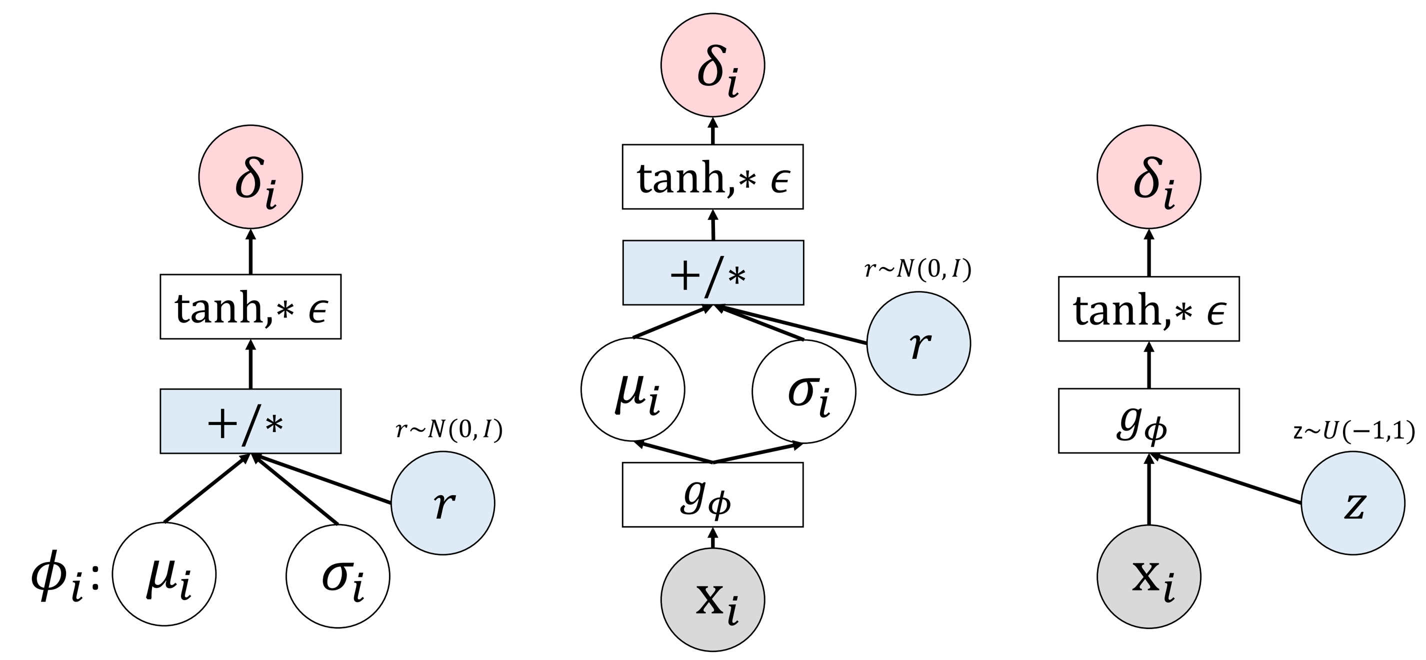
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{p(\delta_i) \in P} J(p(\delta_i), \theta)$$

$$J(p(\delta_i), \theta) = \mathbb{E}_{p(\delta_i)}[L(f_\theta(x_i + \delta_i), y_i)] + \lambda H(p(\delta_i)).$$

Advantages:

- ADT can characterize diverse adversarial examples, many of which may be generated by different attacks, such that ADT leads to better generalizability across attacks.
- The adversarial distributions in ADT can better explore the space of possible adversarial examples, leading to better robustness performance.

Parameterizing Adversarial Distributions



(a) ADT_{EXP} (b) $\text{ADT}_{\text{EXP-AM}}$ (c) $\text{ADT}_{\text{IMP-AM}}$

$\textbf{ADT}_{\textbf{EXP}}$: modeling adversarial perturbations around an input data using a distribution with an explicit density function:

$$\delta_i = \epsilon \cdot \tanh(u_i),\ u_i = N\left(\mu_i, \mathrm{diag}(\sigma_i^2)\right)$$

To estimate the gradient of $J(p_{\phi_i}(\delta_i), \theta)$ with respect to ϕ_i, we adopt the reparameterization trick as

$$\mathbb{E}_{\mathbf{r}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \nabla_{\phi_i}\Big[\mathcal{L}\big(f_\theta(\mathbf{x}_i + \epsilon \cdot \tanh(\boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \mathbf{r})), y_i\big) - \lambda \log p_{\phi_i}\big(\epsilon \cdot \tanh(\boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \mathbf{r})\big)\Big].$$

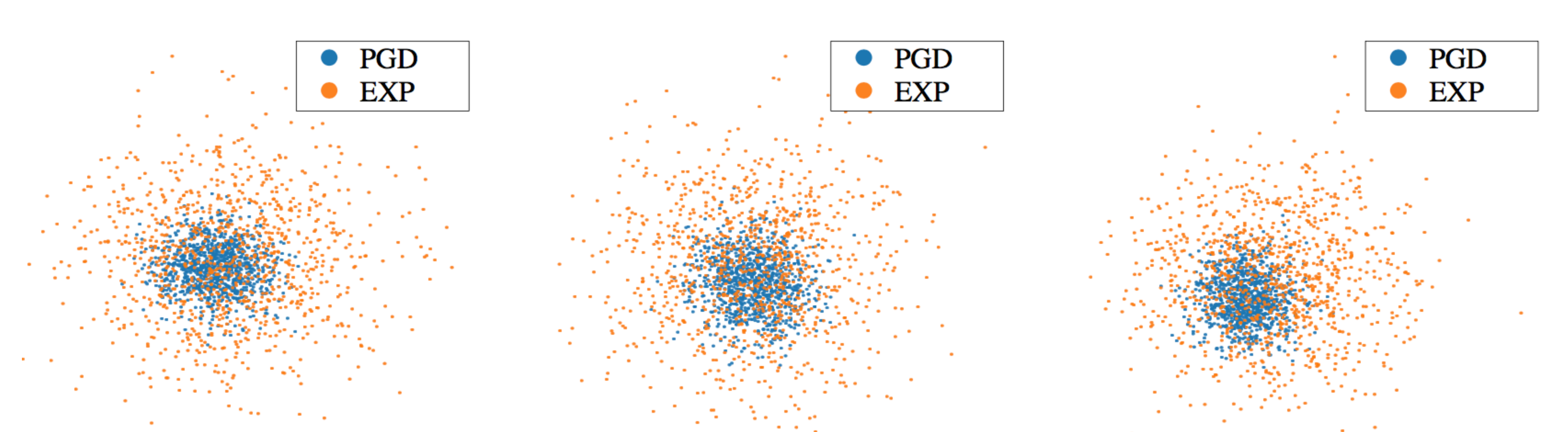
$\textbf{ADT}_{\textbf{EXP-AM}}$: amortizing the inner optimization of ADT_{EXP} by using a conditional generator network. We learn a generator g_ϕ that takes a natural example x_i as input, and outputs the parameters $\{\mu_i, \sigma_i\}$ of the adversarial distribution.

$\textbf{ADT}_{\textbf{IMP-AM}}$: using implicit distributions to characterize the adversarial perturbations. We implicitly define a conditional adversarial distribution as

$$\delta_i = g_\phi(z, x_i), \qquad z{\sim}U(-1,1).$$

Since the entropy of the implicit distributions cannot be estimated exactly, We instead maximize the variational lower bound of the entropy.

Diversity of Adversarial Examples



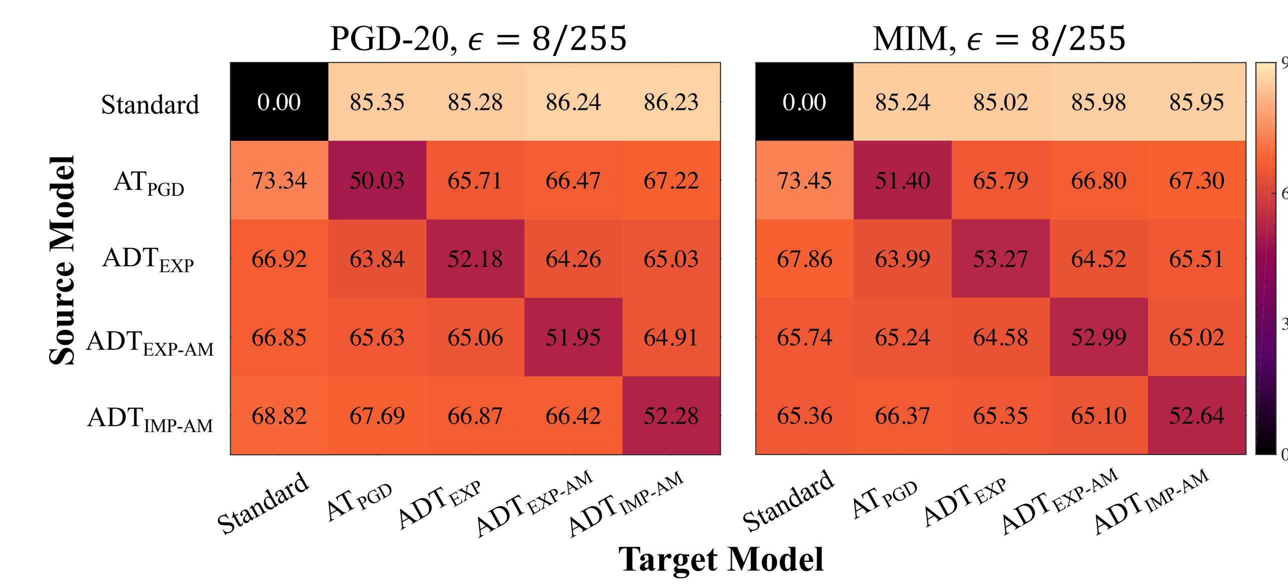
Experiments

Settings: CIFAR-10 with Wide-ResNet-28-10, $\epsilon = 8/255$

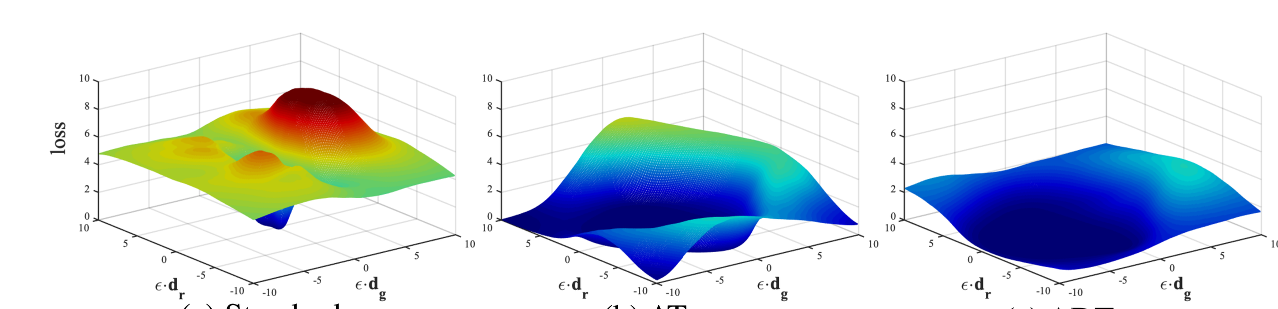
White-box robustness

Model	$\mathcal{A}_{\text{nat}}$	FGSM	PGD-20	PGD-100	MIM	C&W	FeaAttack	$\mathcal{A}_{\text{rob}}$
Standard	**94.81%**	12.05%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AT_{FGSM}	93.80%	79.86%	0.12%	0.04%	0.06%	0.13%	0.01%	0.01%
$\text{AT}_{\text{PGD}}{}^{\dagger}$	87.25%	56.04%	45.88%	45.33%	47.15%	46.67%	46.01%	44.89%
AT_{PGD}	86.91%	58.30%	50.03%	49.40%	51.40%	50.23%	50.46%	48.26%
ALP	86.81%	56.83%	48.97%	48.60%	50.13%	49.10%	48.51%	47.90%
FeaScatter	89.98%	77.40%	70.85%	68.81%	72.74%	58.46%	37.45%	37.40%
ADT_{EXP}	86.89%	60.41%	52.18%	**51.69%**	**53.27%**	52.49%	**52.38%**	**50.56%**
$\text{ADT}_{\text{EXP-AM}}$	87.82%	62.42%	51.95%	51.26%	52.99%	51.75%	52.04%	**50.04%**
$\text{ADT}_{\text{IMP-AM}}$	88.00%	**64.89%**	**52.28%**	51.23%	52.64%	**52.65%**	51.89%	**49.81%**

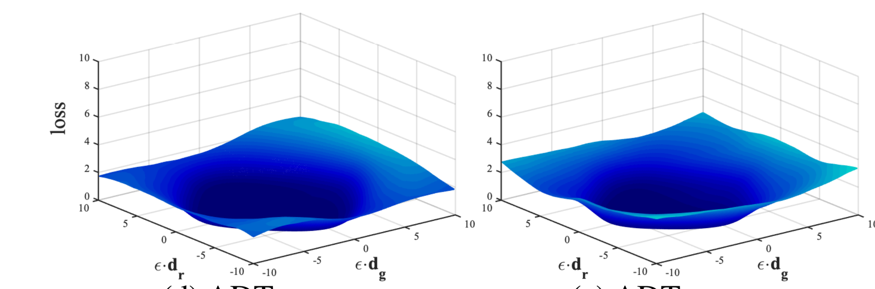
Black-box robustness



Loss landscape visualization



(a) Standard (b) AT_{PGD} (c) ADT_{EXP}



(d) $\text{ADT}_{\text{EXP-AM}}$ (e) $\text{ADT}_{\text{IMP-AM}}$

Model	Dominant eigenvalue
Standard	1.8301±6.3663
AT_{PGD}	0.0242±0.0478
ADT_{EXP}	**0.0180**±0.0311
$\text{ADT}_{\text{EXP-AM}}$	0.0181±**0.0270**
$\text{ADT}_{\text{IMP-AM}}$	0.0211±0.0353

(f) Comparison on the Hessian

Conclusion

- We proposed adversarial distribution training (ADT) framework for learning robust models.
- We introduced three ways to parameterize the adversarial distributions
- We performed extensive experiments to validate the effectiveness of our proposed methods.
- Our code is available at: