Consistency Regularization for Adversarial Robustness

Jihoon Tack[1], Sihyun Yu[1], Jongheon Jeong[1], Minseon Kim[1], Sung Ju Hwang[1,2], Jinwoo Shin[1]

Korea Advanced Institute of Science and Technology (KAIST)[1], AITRICS[2]

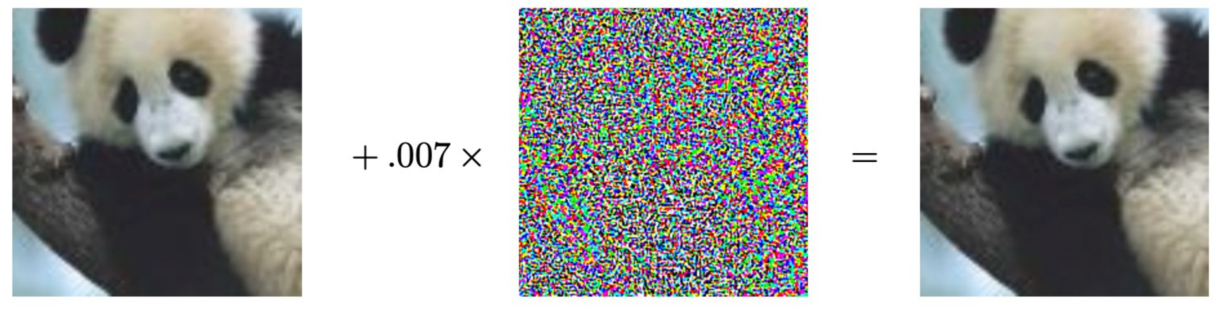
arXiv

Github

TL;DR. We propose an effective consistency regularization technique that prevents robust overfitting by forcing the distribution of attacked augmentations from the same input to be similar

Introduction

Deep neural networks (DNNs) are vulnerable to adversarial noises [1]



$f_\theta(x)$ δ $f_\theta(x+\delta)$

Goal: Train a DNN that is robust to such noise

$$f_\theta(x) = f_\theta(x+\delta), \quad \boxed{\forall\delta} : \|\delta\|_p < \epsilon$$

a classifier

The hardest part

Adversarial training (AT): directly use adversarial examples for training

- Most promising ways to obtain adversarial robustness

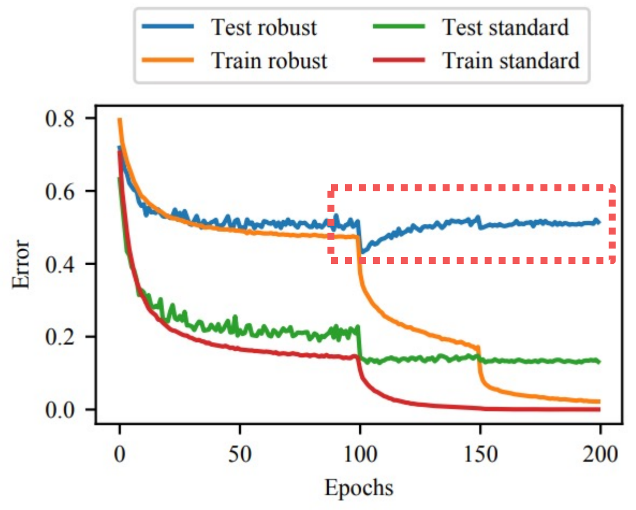
$$\mathcal{L}_{\mathtt{AT}} := \max_{\|\delta\|_p \le \epsilon} \mathcal{L}_{\mathtt{CE}}\big(f_\theta(x+\delta), y\big)$$

One of the most basic form of AT [2]

Robust Overfitting

Problem. AT suffers from robust overfitting [3]

- Test robust error gradually increases from the middle of the training



Only recently, advanced but **sophisticated** training schemes were proposed

🤔 Are there any simpler and more intuitive approaches?

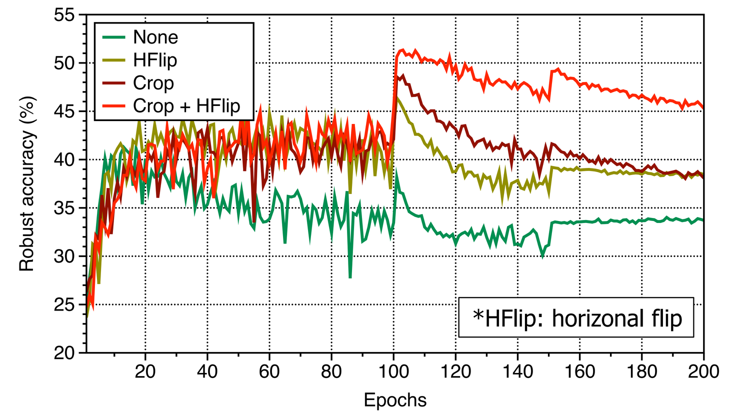
[1] Goodfellow et al. "Explaining and Harnessing Adversarial Examples". ICLR 2015.
[2] Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". ICLR 2018.
[3] Rice et al. "Overfitting in adversarially robust deep learning". ICML 2020.

Consistency Regularization for AT

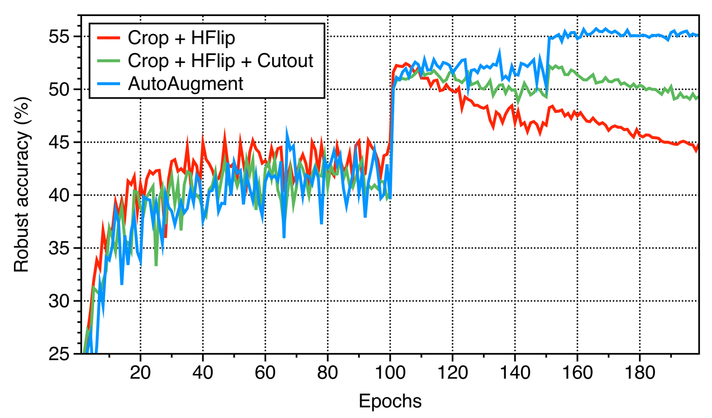
Data augmentations (DAs) can somewhat reduce the robust overfitting.

$$\max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\mathtt{CE}}\Big(f_\theta\big(T(x)+\delta\big), y\Big) \quad \text{where} \quad T \sim \underbrace{\mathcal{T}_{\mathtt{conven}}}_{\text{crop, flip}} \cup \underbrace{\mathcal{T}_{\mathtt{add}}}_{\text{+ color, cutout, grayscale}}$$

- Conventional DAs, e.g., crop, is already useful for reducing overfitting.
- Additional DAs, e.g., color jitter, is quite effective to reduce overfitting.



Conventional DAs



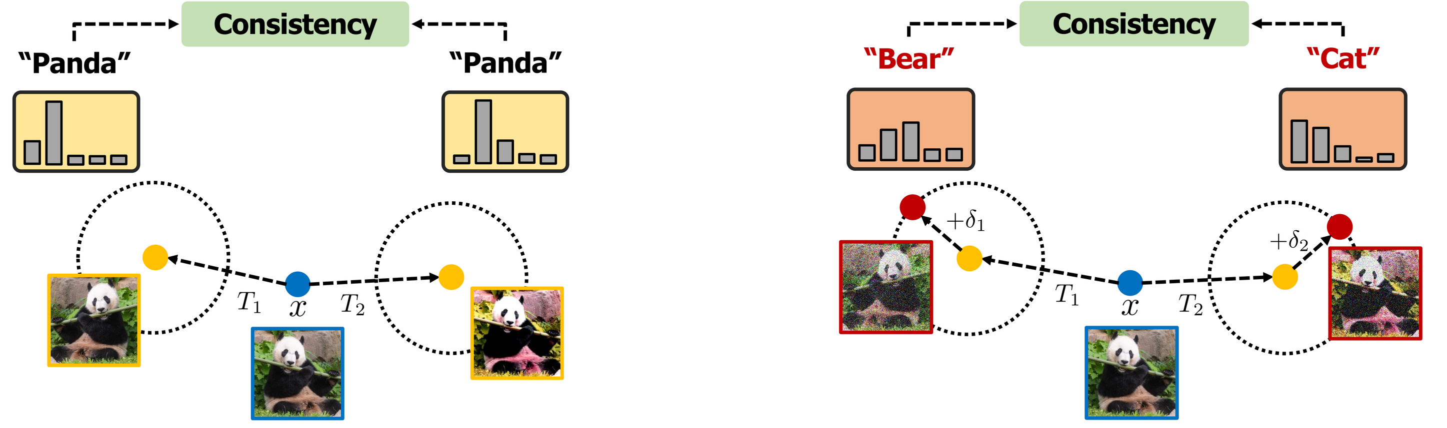
Additional DAs

Consistency regularization (CR) can further improve the robustness

$$\mathtt{JS}\Big(\hat{f}_\theta\big(T_1(x)+\delta_1;\tau\big) \,\|\, \hat{f}_\theta\big(T_2(x)+\delta_2;\tau\big)\Big) \quad \text{where} \quad T_1, T_2 \sim \mathcal{T}$$

$\hat{f}_\theta$: *temperature (τ) scaled classifier*

(+) Easy-to-use: scalable, and hyperparameter-efficient
(+) Flexible: Can be applied to any AT schemes



Conventional CR

Proposed CR

🔍 **Finding. Attack direction** contains intrinsic information.

- Most frequently attacked class is the **most confusing class**

$\mathtt{argmax}_{k \ne y} f_\theta^{(k)}(x)$: top-1 prediction disregarding the true class

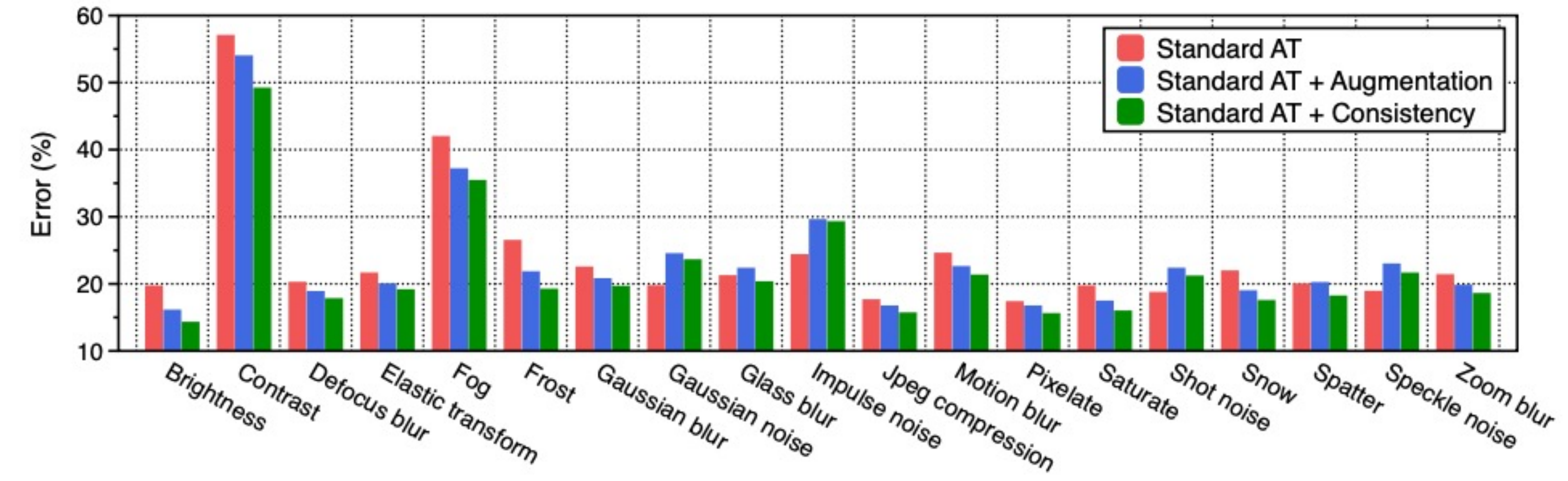
- Matching the attack direction injects a **strong inductive bias!**

Experimental Results

Main results. Our method shows the effectiveness mainly for three parts: (1) reducing overfitting, (2) unseen adversaries, (3) common corruptions

Dataset (Architecture)	Method	Clean	PGD-20	PGD-100	CW_∞	AutoAttack
CIFAR-10 (PreAct-ResNet-18)	Standard (Madry et al. 2018)	84.57 (83.43)	45.04 (52.82)	44.86 (52.67)	44.31 (50.66)	40.43 (47.63)
	+ Consistency	**86.45** (85.25)	**56.51** (57.53)	**56.38** (57.39)	**52.45** (52.70)	**48.57** (49.05)
	TRADES (Zhang et al. 2019)	82.87 (82.13)	50.95 (53.98)	50.83 (53.85)	49.30 (51.71)	46.32 (49.32)
	+ Consistency	**83.63** (83.55)	**55.00** (55.16)	**54.89** (54.98)	**49.91** (50.67)	**47.68** (49.01)
	MART (Wang et al. 2020)	82.63 (77.00)	51.12 (54.83)	50.91 (54.74)	46.92 (49.26)	43.46 (46.74)
	+ Consistency	**83.43** (81.89)	**59.59** (60.48)	**59.52** (60.47)	**51.78** (51.83)	**48.91** (48.95)
CIFAR-10 (WideResNet-34-10)	Standard (Madry et al. 2018)	86.37 (87.55)	50.16 (55.86)	49.80 (55.65)	49.25 (54.45)	45.62 (51.24)
	+ Consistency	**89.82** (89.93)	**58.63** (61.11)	**58.41** (60.99)	**56.38** (57.80)	**52.36** (54.08)
	TRADES (Zhang et al. 2019)	85.05 (84.30)	51.20 (57.34)	50.89 (57.20)	50.88 (55.08)	46.17 (53.02)
	+ Consistency	**87.71** (87.92)	**58.39** (59.12)	**58.19** (58.99)	**54.84** (55.97)	**51.94** (53.11)
	MART (Wang et al. 2020)	85.75 (83.98)	49.31 (57.28)	49.06 (57.22)	48.05 (53.21)	44.96 (50.62)
	+ Consistency	**87.17** (85.81)	**63.26** (64.95)	**62.81** (64.80)	**57.46** (56.24)	**52.41** (53.33)
CIFAR-100 (PreAct-ResNet-18)	Standard (Madry et al. 2018)	57.13 (57.10)	22.36 (29.67)	22.25 (29.65)	21.97 (27.99)	19.85 (25.38)
	+ Consistency	**62.73** (61.62)	**30.75** (32.33)	**30.62** (32.24)	**27.63** (28.39)	**24.55** (25.52)
Tiny-ImageNet (PreAct-ResNet-18)	Standard (Madry et al. 2018)	41.54 (45.26)	11.71 (20.92)	11.60 (20.87)	11.20 (18.72)	9.63 (16.03)
	+ Consistency	**50.15** (49.46)	**21.33** (23.31)	**21.24** (23.24)	**19.08** (20.29)	**15.69** (16.90)

		l_∞		l_2		l_1	
Dataset	Method \ ϵ	4/255	16/255	150/255	300/255	2000/255	4000/255
CIFAR-10	Standard (Madry et al. 2018)	65.93	19.23	52.56	25.68	45.96	36.85
	+ Consistency	**73.74**	**23.47**	**65.81**	**36.87**	**58.66**	**50.79**
	TRADES (Zhang et al. 2019)	68.30	24.17	56.14	28.94	44.08	29.58
	+ Consistency	**70.33**	**26.52**	**63.70**	**39.16**	**56.48**	**48.32**
	MART (Wang et al. 2020)	67.76	23.36	57.17	30.98	46.61	34.63
	+ Consistency	**72.67**	**30.31**	**66.17**	**43.76**	**60.57**	**54.19**
CIFAR-100	Standard (Madry et al. 2018)	36.14	7.37	27.97	11.98	30.48	27.29
	+ Consistency	**46.11**	**11.53**	**39.77**	**20.69**	**36.04**	**32.75**
Tiny-ImageNet	Standard (Madry et al. 2018)	23.23	2.69	28.05	17.80	33.30	31.55
	+ Consistency	**34.18**	**5.74**	**40.06**	**30.62**	**43.90**	**42.65**



Analysis on attack directions.

- 77.45% out of the misclassified adversarial examples predicts the most confusing class of 'clean' input.
- i.e., most confident prediction expect for the true class