Consistency Regularization for Adversarial Robustness

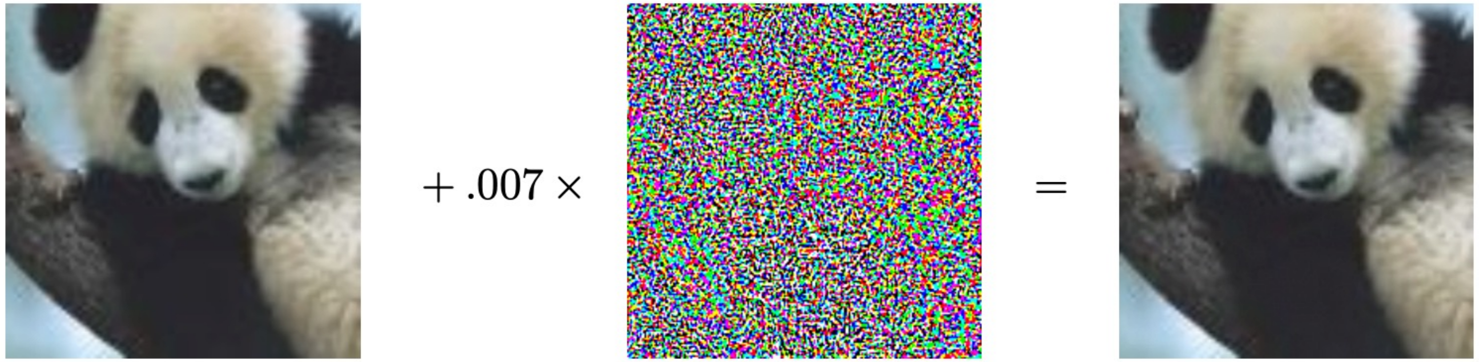
Jihoon Tack[1], Sihyun Yu[1], Jongheon Jeong[1], Minseon Kim[1], Sung Ju Hwang[1,2], Jinwoo Shin[1]

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

AAAI Conference on Artificial Intelligence 2022

Adversarial Examples in DNNs

Deep neural networks (DNNs) are vulnerable to adversarial noises



$f_\theta(x)$ $\quad\quad$ δ $\quad\quad$ $f_\theta(x+\delta)$

Fundamental question: Can we train DNNs that are robust to such noises?

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

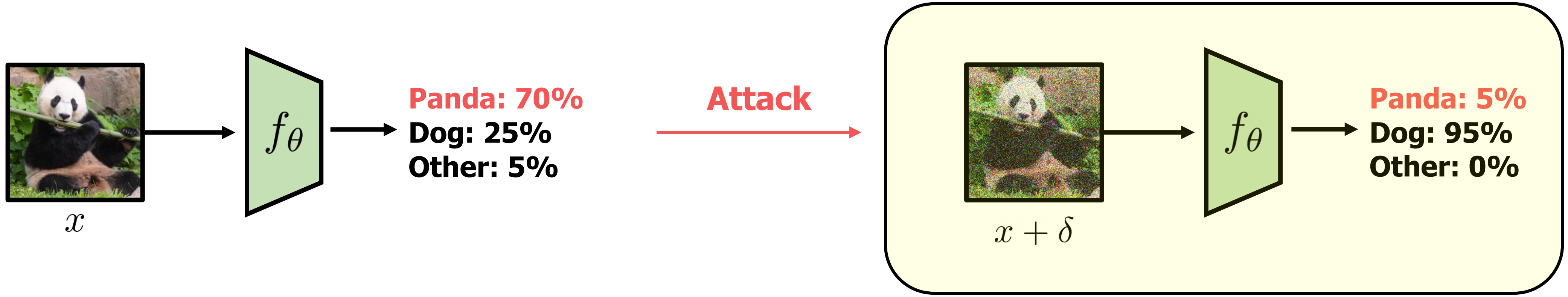
a classifier

The hardest part

[Goodfellow et al., ICLR 2015] Explaining and Harnessing Adversarial Examples.

Adversarial Training

Adversarial Training (AT) directly incorporate adversarial examples for training



Use adversarial example to train the network

- Madry et al., 2018: generate adversarial example during training via min-max optimization

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

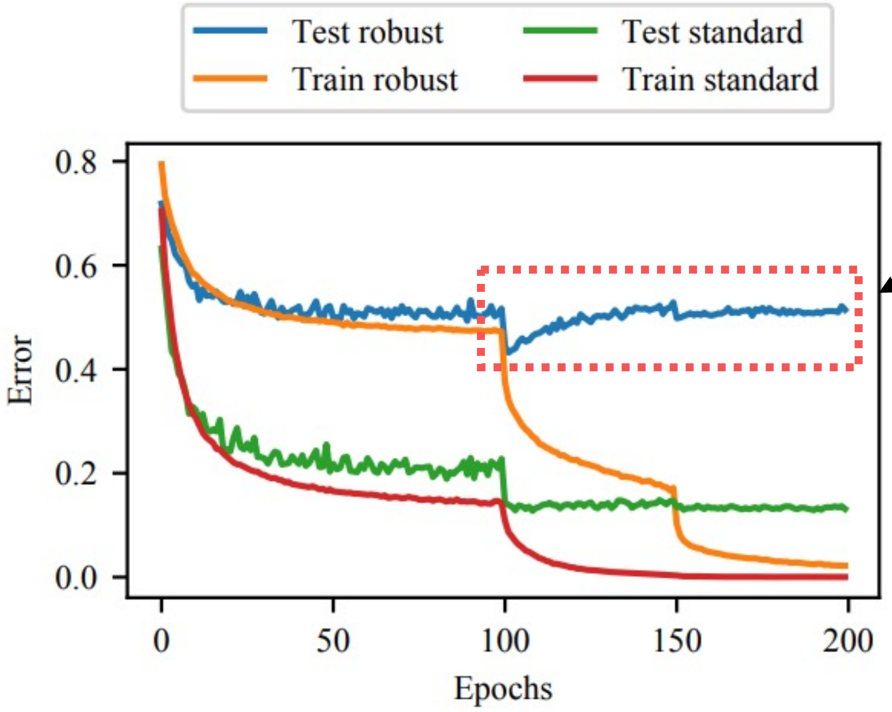
One of the most basic form of AT

[Madry et al., ICLR 2018] Towards Deep Learning Models Resistant to Adversarial Attacks

Robust Overfitting [Rice et al., ICML 2020]

Problem: AT suffers from robust overfitting

- The robust error of test set, gradually increases from the middle of training
- Make practitioners consider a bag of tricks for a successful training, e.g., early stopping



overfitting occurs

[Rice et al., ICML 2020] Overfitting in adversarially robust deep learning.

Robust Overfitting [Rice et al., ICML 2020]

Problem: AT suffers from robust overfitting

- The robust error of test set, gradually increases from the middle of training
- Make practitioners consider a bag of tricks for a successful training, e.g., early stopping

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

- E.g., adversarial weight perturbation (Wu et al., 2020), self-training (Chen et al., 2021)

Is there a simpler and more intuitive approach?

[Rice et al., ICML 2020] Overfitting in adversarially robust deep learning.
[We et al., NeurIPS 2020] Adversarial Weight Perturbation Helps Robust Generalization.
[Chen et al., ICLR 2021] Robust Overfitting may be mitigated by properly learned smoothening

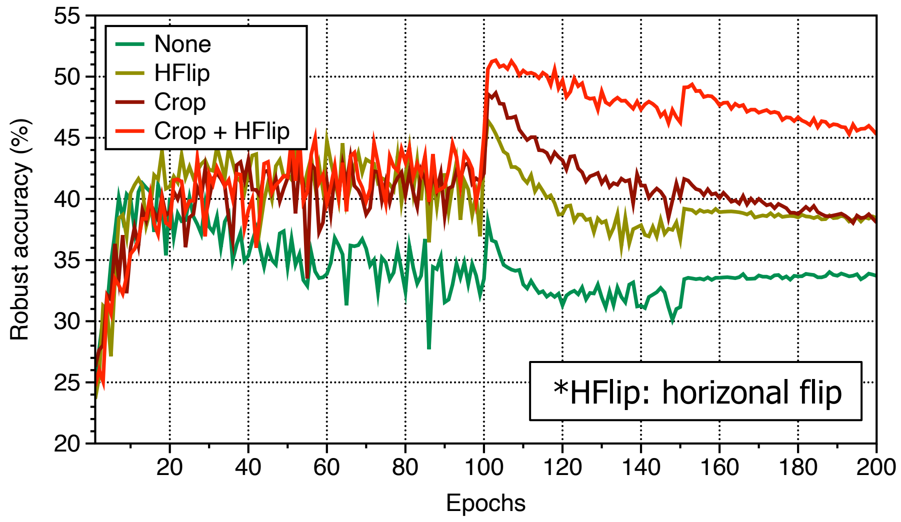
Data Augmentations can reduce Overfitting

We found that data augmentations (DAs) is important for robust overfitting

$$\max_{||\delta||_\infty \leq \epsilon} \mathcal{L}_{\texttt{CE}}\Big(f_\theta\big(T(x)+\delta\big), y\Big) \quad \text{where} \quad T \sim \mathcal{T}_{\texttt{conven}}$$

random cropping, horizonal flip

- 1) Conventional DAs, e.g., cropping, is already somewhat useful for reducing robust overfitting



1) Conventional DAs

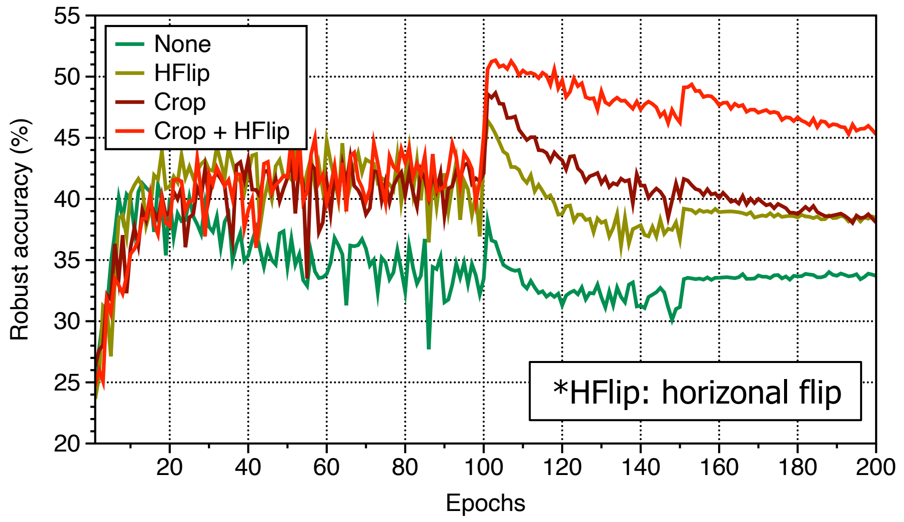
Data Augmentations can reduce Overfitting

We found that data augmentations (DAs) is important for robust overfitting

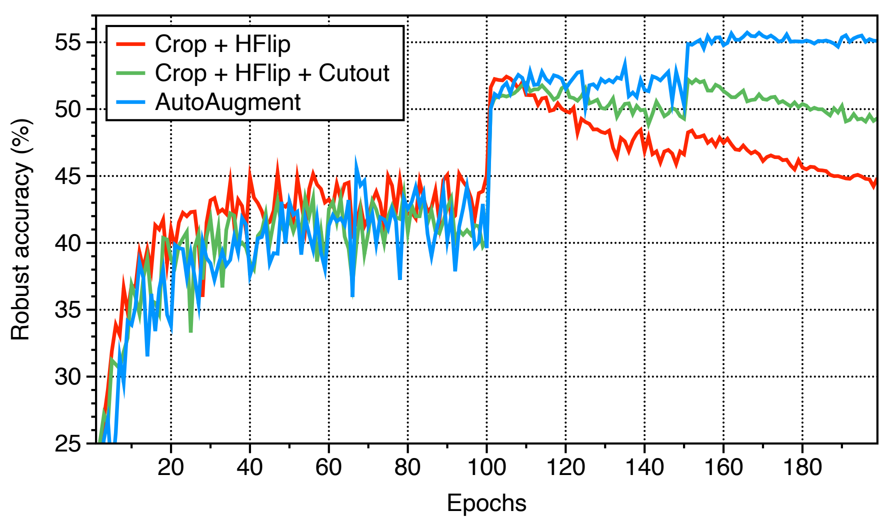
$$\max_{||\delta||_\infty \le \epsilon} \mathcal{L}_{\texttt{CE}}\Big(f_\theta\big(T(x)+\delta\big), y\Big) \quad \text{where} \quad T \sim \mathcal{T}_{\texttt{conven}} \cup \boxed{\mathcal{T}_{\texttt{add}}}$$

+ AutoAugment

- 1) Conventional DAs, e.g., cropping, is already somewhat useful for reducing robust overfitting
- 2) Additional DAs to conventional choices, e.g., AutoAugment, is effective to reduce overfitting



1) Conventional DAs



2) Additional DAs

Consistency Regularization for AT

Consistency regularization (CR) can further improve robust generalization!

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

temperature (τ) scaled classifier

independently sampled augmentation

- The proposed scheme is easy-to-use, and flexible (can be applied to various AT schemes)

Consistency Regularization for AT

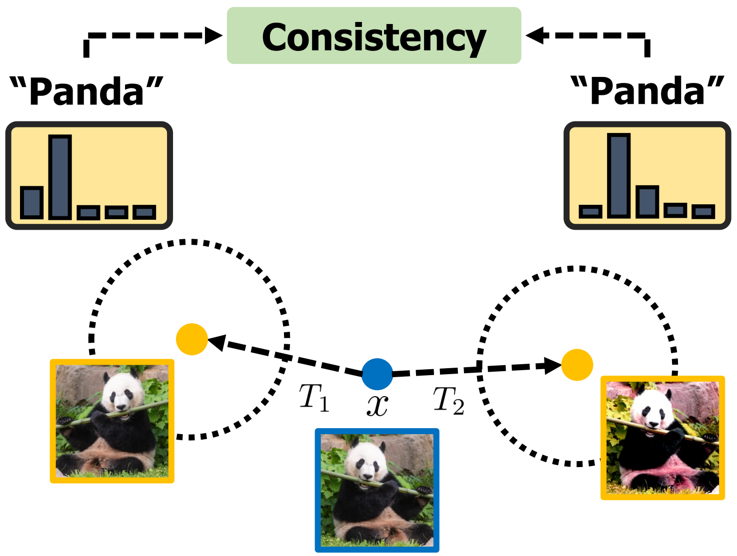
Consistency regularization (CR) can further improve robust generalization!

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

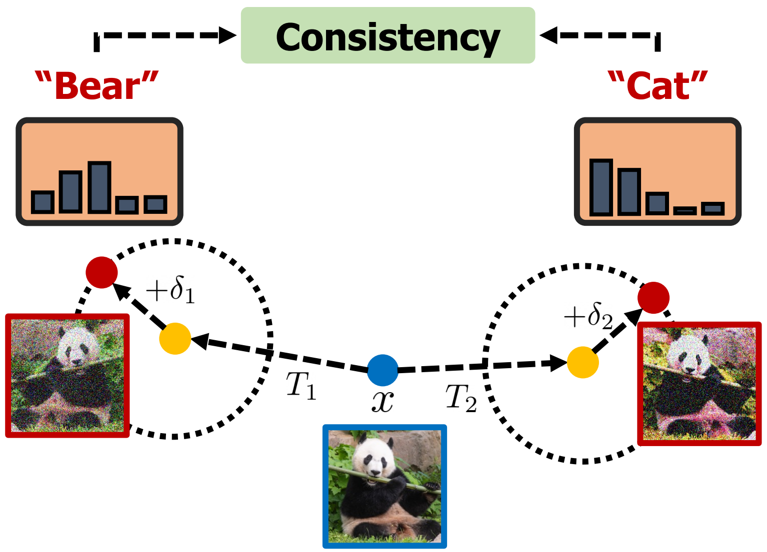
temperature (τ) scaled classifier

independently sampled augmentation

- The proposed scheme is easy-to-use, and flexible (can be applied to various AT schemes)



Conventional CR



Proposed CR

Consistency Regularization for AT

Consistency regularization (CR) can further improve robust generalization!

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

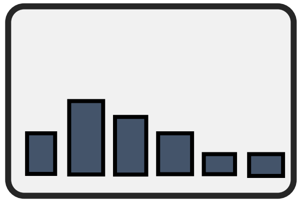
temperature (τ) scaled classifier

independently sampled augmentation

- The proposed scheme is easy-to-use, and flexible (can be applied to various AT schemes)

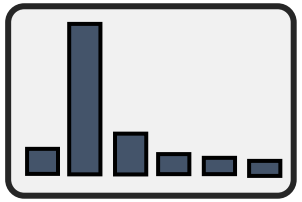
$$\hat{f}^c_\theta(x;\tau)=\frac{\exp(z_c/\tau)}{\sum_{i\in\mathcal{C}}\exp(z_i/\tau)}$$

τ : temperature
z_i: logit of class i



τ > 1

Use small τ to sharpen the distribution



τ < 1

Consistency Regularization for AT

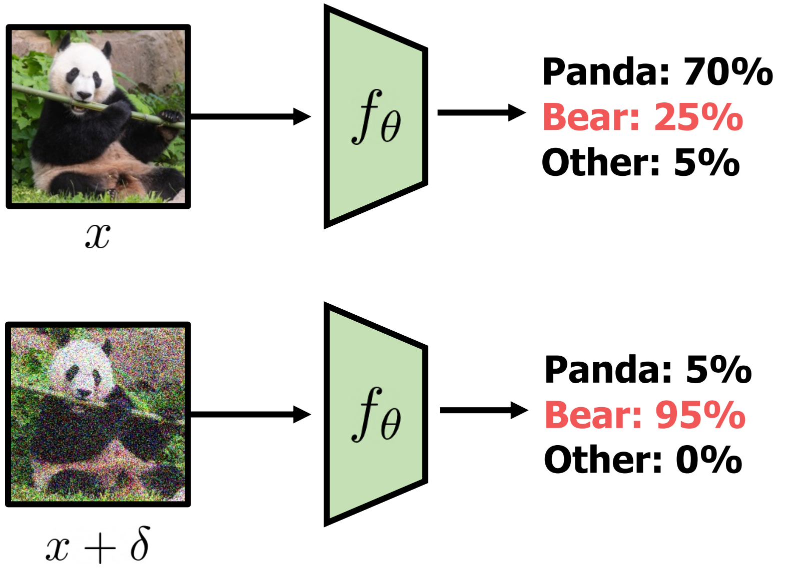
Consistency regularization (CR) can further improve robust generalization!

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

temperature (τ) scaled classifier

independently sampled augmentation

- The proposed scheme is easy-to-use, and flexible (can be applied to various AT schemes)



Attack direction itself contains intrinsic information

- Most frequently attacked class is the most confusing class

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

- Matching the attack direction injects a strong inductive bias!

Consistency Regularization for AT

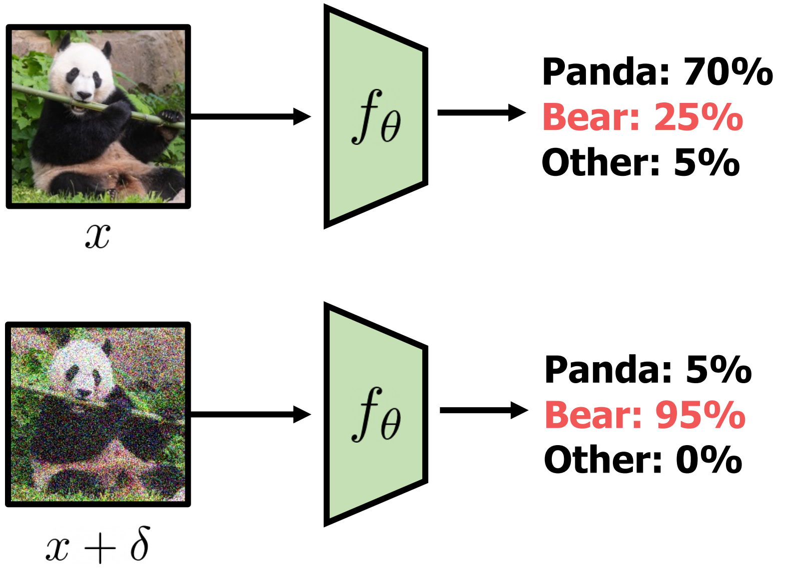
Consistency regularization (CR) can further improve robust generalization!

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

temperature (τ) scaled classifier

independently sampled augmentation

- The proposed scheme is easy-to-use, and flexible (can be applied to various AT schemes)



Attack direction consistency is important

- Utilizing conventional consistency can degrade the accuracy

Loss	Clean	PGD-100
AT (3)	85.41	55.18
AT (3) + previous CR (5)	88.01	53.11
AT (3) + proposed CR (4)	86.45	56.38

Experimental Results

Consistency regularization demonstrates the effectiveness mainly for three parts

- 1) Reduce robust overfitting (+ improves robustness also)

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)

Experimental Results

Consistency regularization demonstrates the effectiveness mainly for three parts

- 2) Robust against unseen adversaries [Tramer et al., 2019]

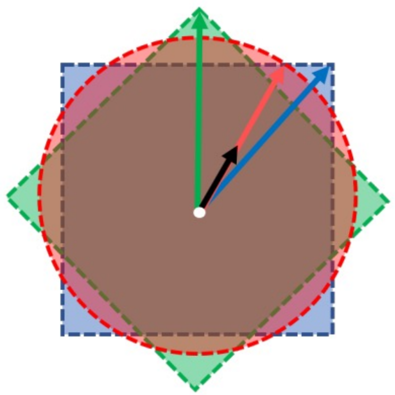


Figure 1: A depiction of the steepest descent directions for ℓ_∞, ℓ_2, and ℓ_1 norms. The gradient is the black arrow, and the α radius step sizes and their corresponding steepest descent directions ℓ_∞, ℓ_2, and ℓ_1 are shown in blue, red, and green respectively.

Unseen adversaries are hard to defense

- We train the model on l_∞ perturbation and test on l_1, l_2
- We also test different attack radii of ϵ

[Tramer et al., NeurIPS 2019] Adversarial training and robustness for multiple perturbations.
[Maini et al., ICML 2020] Adversarial Robustness Against the Union of Multiple Perturbation Models

Experimental Results

Consistency regularization demonstrates the effectiveness mainly for three parts

- 2) Robust against unseen adversaries [Tramer et al., 2019]

Dataset	Method \ ϵ	l_∞ 4/255	l_∞ 16/255	l_2 150/255	l_2 300/255	l_1 2000/255	l_1 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**

[Tramer et al., NeurIPS 2019] Adversarial training and robustness for multiple perturbations.
[Maini et al., ICML 2020] Adversarial Robustness Against the Union of Multiple Perturbation Models

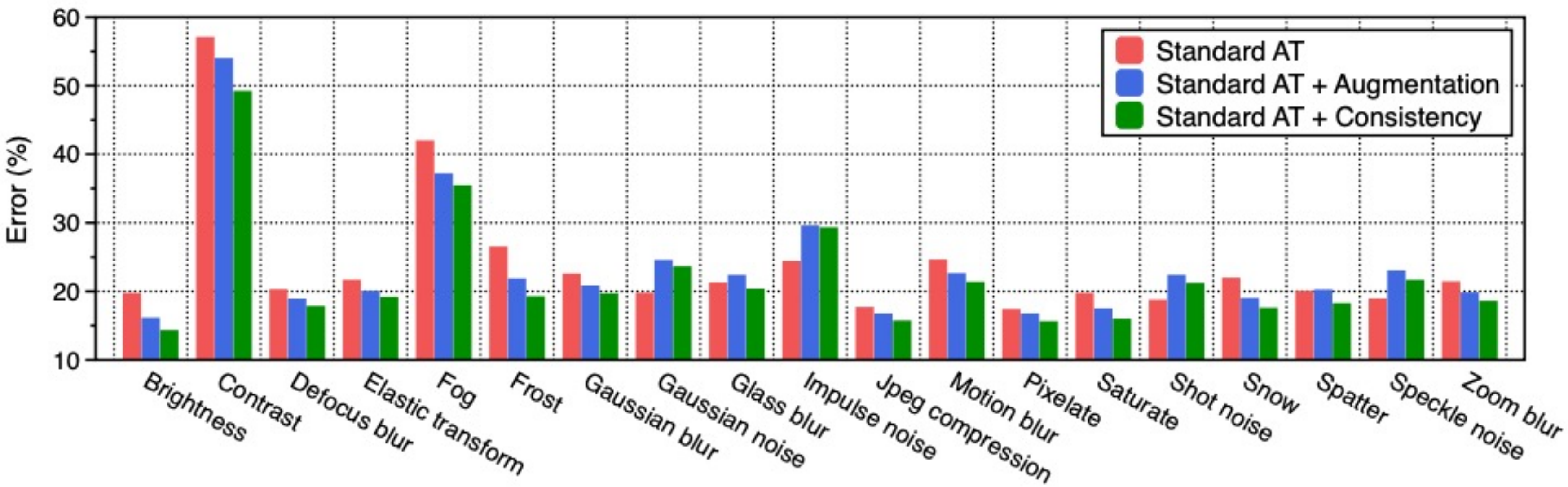
Experimental Results

Consistency regularization demonstrates the effectiveness mainly for three parts

- 3) Robust against common corruptions [Hendrycks et al., 2019]

Method	mCE ↓
Standard cross-entropy	27.02
Standard (Madry et al. 2018) **+ Consistency**	24.03 **21.83**
TRADES (Zhang et al. 2019) **+ Consistency**	25.50 **23.95**
MART (Wang et al. 2020) **+ Consistency**	26.20 **24.41**

Mean corruption error (mCE) of PreAct-ResNet-18 trained on CIFAR-10.



Classification error (%) on each corruption type of CIFAR-10-C

[Hendrycks et al., ICLR 2019] Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

Experimental Results

Consistency regularization demonstrates the effectiveness mainly for three parts

- Our method method somewhat surpass the performance of the recent regularization technique

Dataset	Method	Clean	l_∞ (Seen)			l_2 (Unseen)		l_1 (Unseen)	
			PGD-100 (8/255)	CW_∞ (8/255)	AutoAttack (8/255)	PGD-100 (150/255)	PGD-100 (300/255)	PGD-100 (2000/255)	PGD-100 (4000/255)
CIFAR-10	Standard (Madry et al. 2018)	84.57	44.86	44.31	40.43	52.56	25.68	45.96	36.85
	+ AWP (Wu, Xia, and Wang 2020)	80.34	55.39	52.31	**49.60**	61.39	36.05	56.30	48.37
	+ Consistency	**86.45**	**56.38**	**52.45**	48.57	**65.81**	**36.87**	**58.66**	**50.79**
CIFAR-100	Standard (Madry et al. 2018)	56.96	20.86	21.20	18.93	27.65	11.08	26.49	21.48
	+ AWP (Wu, Xia, and Wang 2020)	52.91	30.06	26.42	24.32	35.71	20.18	33.63	30.38
	+ Consistency	**62.73**	**30.62**	**27.63**	**24.55**	**39.77**	**20.69**	**36.04**	**32.75**
Tiny-ImageNet	Standard (Madry et al. 2018)	41.54	11.60	11.20	9.63	28.05	17.80	33.30	31.55
	+ AWP (Wu, Xia, and Wang 2020)	40.25	20.64	18.05	15.26	33.31	26.86	35.48	34.22
	+ Consistency	**50.15**	**21.24**	**19.08**	**15.69**	**40.06**	**30.62**	**43.90**	**42.65**

[Wu et al., NeurIPS 2020] Adversarial Weight Perturbation Helps Robust Generalization

Ablation Study

We verify the effectiveness of each component

- (a) data augmentation, (b) consistency regularization loss
- The performance improves step by step with the addition of the component

Method	PGD-100	mCE ↓
Standard (Madry et al. 2018)	44.86	24.03
+ Cutout (DeVries and Taylor 2017)	49.95	24.05
+ AutoAugment (Cubuk et al. 2019)	55.18	23.38
+ Consistency	**56.38**	**22.06**

We also verify the effectiveness of the temperature scaling

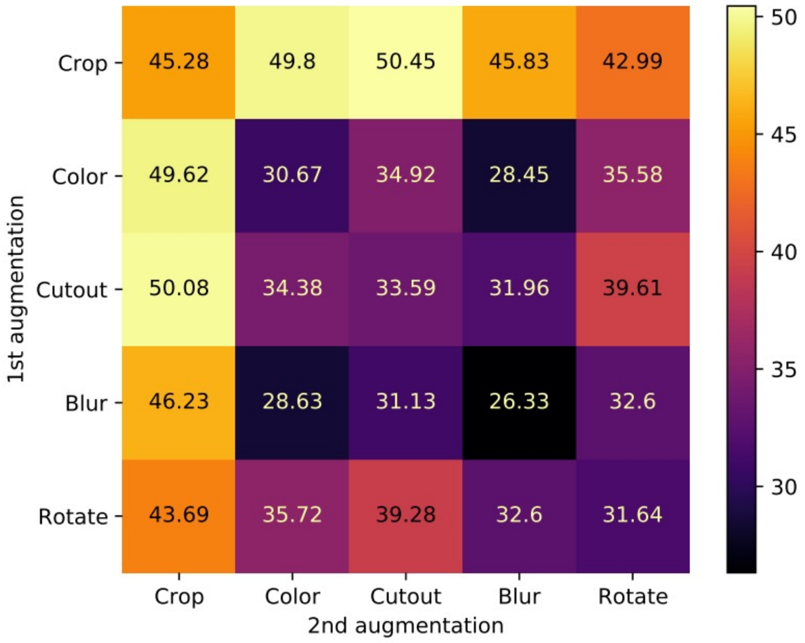
- As our intuition, sharpening the prediction with small temperature shows an improvement

τ	0.5	0.8	1.0	2.0	5.0
PGD-100	**56.38**	56.22	55.79	56.04	55.57

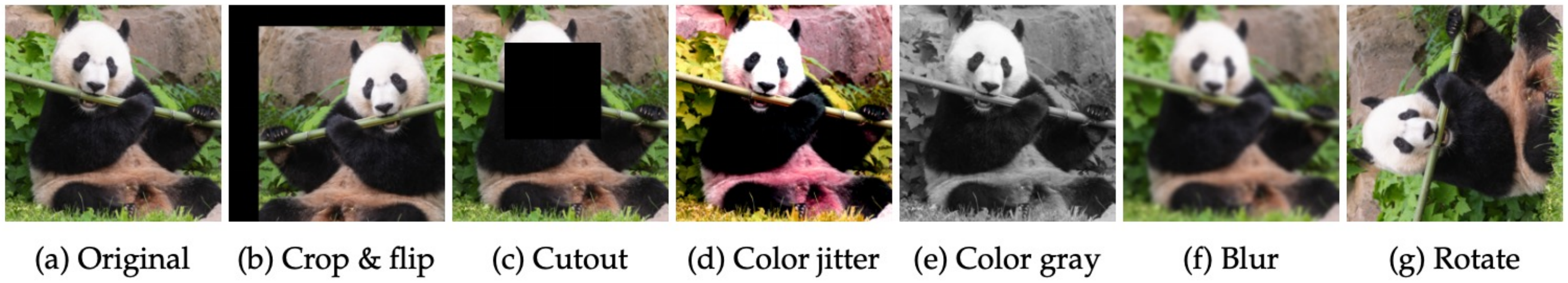
Analysis on Data Augmentations

Which augmentation family improve the generalization in adversarial training?

- We observe that cropping, Cutout and color transformation shows effectiveness
- We hypothesize that sample diversity through augmentations is significant for the improvement



PGD-100 accuracy (%)
under the composition of augmentations



Visualization of augmentations

Take-home message

Data augmentation is quite effective for preventing the robust overfitting

Consistency regularization can further improve the robustness

- However, one should match the attack direction to be consistent

Our method can improve robustness of

- (1) seen adversaries, (2) unseen adversaries, and (3) natural corruptions

Thank you for your attention ☺

Paper: https://arxiv.org/abs/2103.04623

Code: https://github.com/alinlab/consistency-adversarial