JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

#THU-PM-388

Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations

Lei Hsiung[1,4], Yun-Yun Tsai[2], Pin-Yu Chen[3], Tsung-Yi Ho[1,4]

[1]National Tsing Hua University [2]Columbia University [3]IBM Research [4]CUHK

國立清華大學 NATIONAL TSING HUA UNIVERSITY

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

IBM

香港中文大學 The Chinese University of Hong Kong

Project Page

Project Page

Demonstration

Standard

ℓ_∞-robust

GAT

(Proposed)

Hue

-0.4π -0.2π Normal 0.2π 0.4π

Saturation

50% 75% Normal 150% 200%

Rotation

-10˚ -5˚ Normal 5˚ 10˚

Brightness

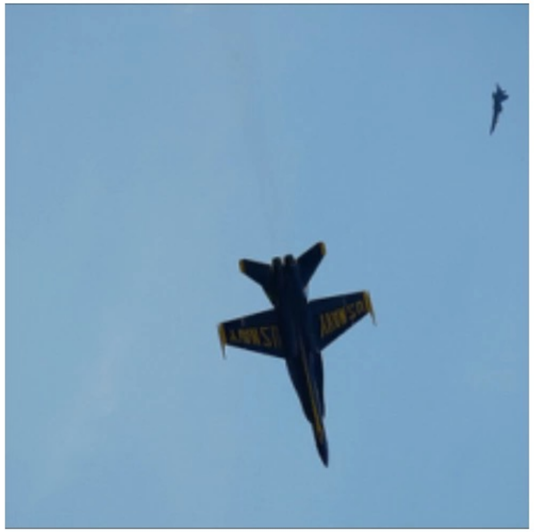
50% 75% Normal 150% 200%

Contrast

50% 75% Normal 150% 200%

PGD (ℓ_∞)

Normal 1/255 2/255 3/255 4/255



Ground Truth: Warplane

Model Prediction: **Warplane**

Reset

Project Page

Demonstration

Standard

ℓ_∞-robust

GAT

(Proposed)

Hue

-0.4π -0.2π Normal 0.2π 0.4π

Saturation

50% 75% Normal 150% 200%

Rotation

-10° -5° Normal 5° 10°

Brightness

50% 75% Normal 150% 200%

Contrast

50% 75% Normal 150% 200%

PGD (ℓ_∞)

Normal1/255 2/255 3/255 4/255

Ground Truth: Warplane

Model Prediction: **Starfish**

Reset

Browse on: https://hsiung.cc/CARBEN/

Project Page

Demonstration

Standard

ℓ_∞-robust

GAT

(Proposed)

Hue

-0.4π -0.2π Normal 0.2π 0.4π

Saturation

50% 75% Normal 150% 200%

Rotation

-10° -5° Normal 5° 10°

Brightness

50% 75% Normal 150% 200%

Contrast

50% 75% Normal 150% 200%

PGD (ℓ_∞)

Normal 1/255 2/255 3/255 4/255

Ground Truth: Warplane

Model Prediction: **Warplane**

Reset

Browse on: https://hsiung.cc/CARBEN/

Background

- Deep neural networks (DNNs) have shown remarkable success in many real-life applications. In fact, it is easy for neural networks to achieve excellent performance on benign data points.
- However, recent researches have shown that one can intentionally derive an adversarial example, and make it imperceptible to human beings.

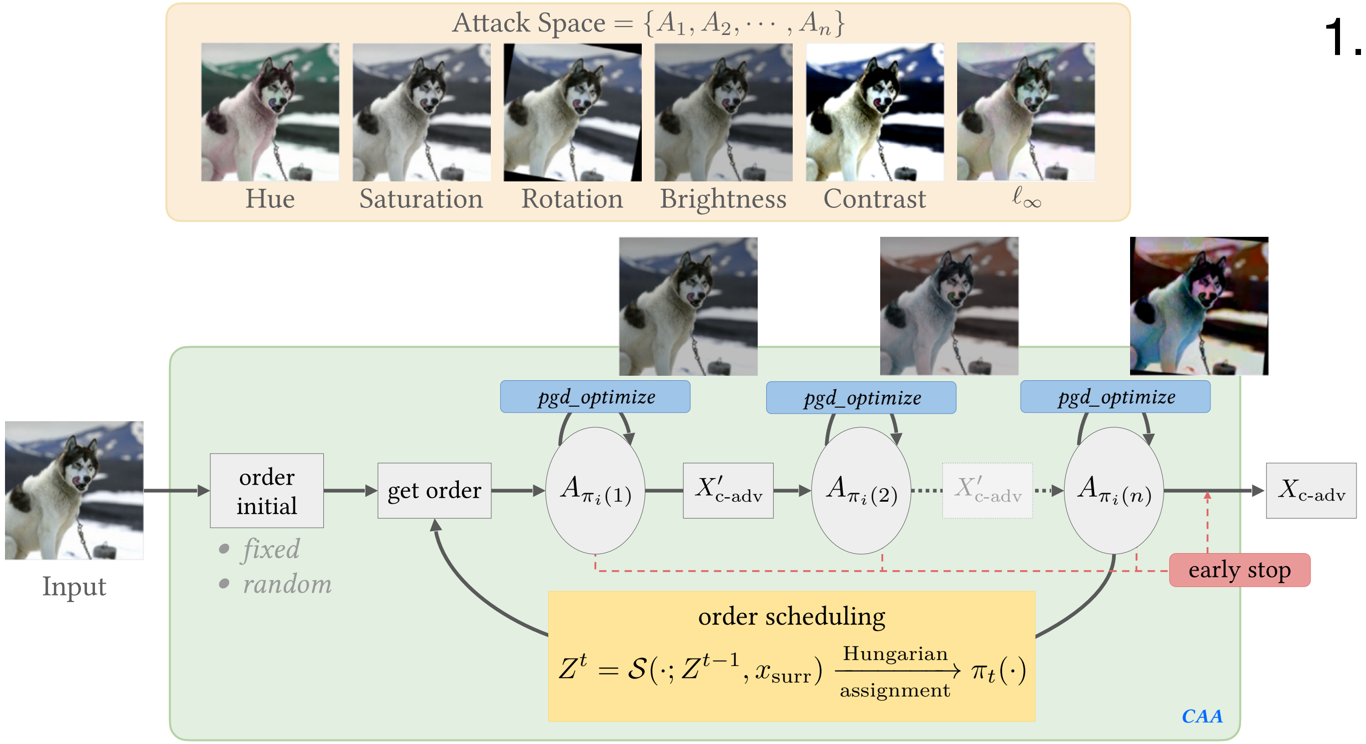
[1] EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples, P.-Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C-.J. Hsieh, AAAI 2018
[2] Explaining and Harnessing Adversarial Examples, I. Goodfellow, J. Shlens, C. Szegedy, ICLR 2015

Motivations

- We propose **Composite Adversarial Attacks (CAA)** to generate hardened adversarial examples and **Generalized Adversarial Training (GAT)** to obtain a robust model against them.

Methodology

Composite Adversarial Attacks (CAA)



1. Generate a Surrogate Image to updating the scheduling matrix Z.

$$Z^\top = [\mathbf{z}_1, \cdots, \mathbf{z}_n]$$

$$x^i_{\text{surr}} = \sum_{j=1}^{n} z_{ij} \cdot A_j(x^{i-1}_{\text{surr}}; \delta_j)), \forall i \in \{1, \ldots, n\}$$

$$\begin{aligned} x^n_{\text{surr}} &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(\mathbf{z}_1^\top \mathbf{A}(x)))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(\sum_{j=1}^{n} z_{1j} \cdot A_j(x; \delta_j)))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(x^1_{\text{surr}}))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(x^2_{\text{surr}})) = x_{\text{surr}} \end{aligned}$$

2. Derive the Optimal Attack Order by updating the scheduling matrix Z^t.

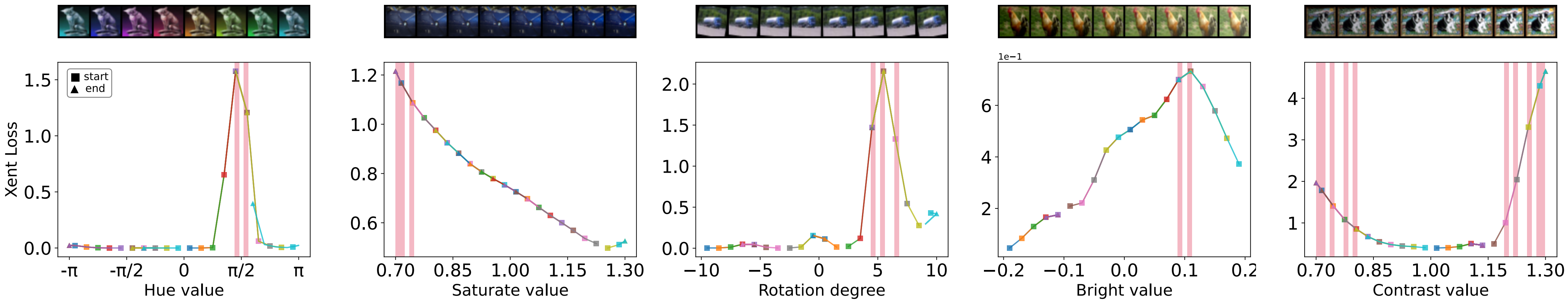
$$Z^t = \mathcal{S}\big(\exp(Z^{t-1} + \frac{\partial \mathcal{L}(\mathcal{F}(x_{\text{surr}}), y)}{\partial Z^{t-1}})\big), \ \mathcal{S}: \text{Sinkhorn normalization.}$$

Methodology

Composite Adversarial Attacks (CAA)

- The Component-wise PGD
  - δ_k: the perturbation value of the semantic attack A_k
  - α : the step size of the updating process

$$\delta_k^{t+1} = \mathrm{clip}_{\epsilon_k}\left(\delta_k^t + \alpha \cdot \mathrm{sign}(\nabla_{\delta_k^t} \mathcal{L}(\mathcal{F}(A_k(x;\delta_k^t)), y))\right)$$



Demonstration

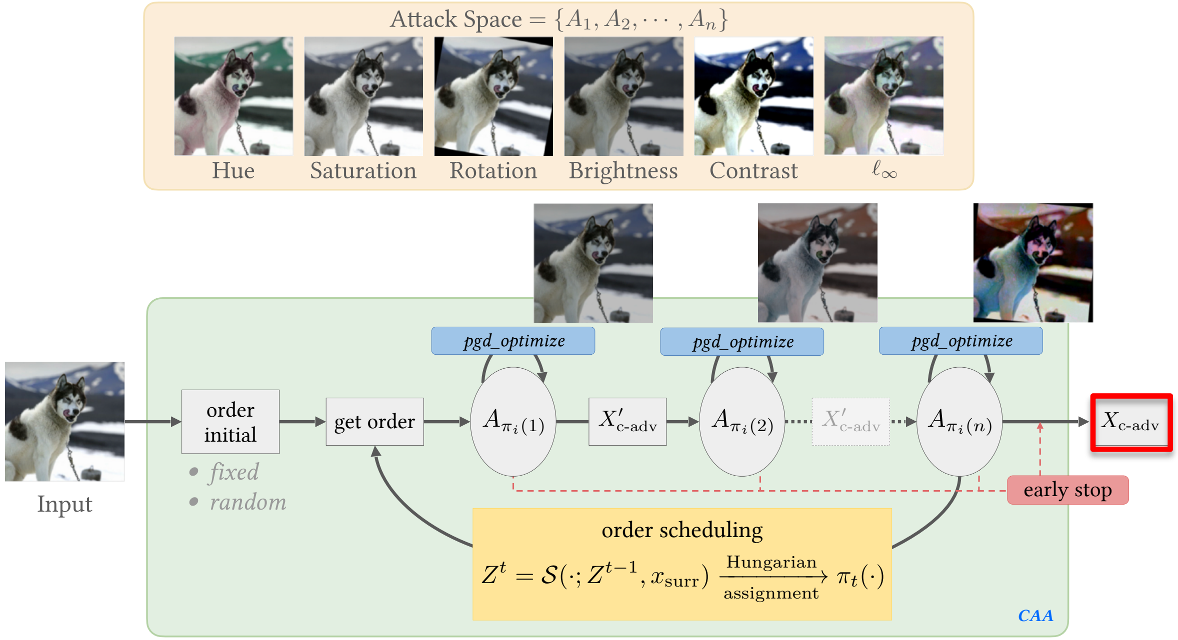
- Composite Adversarial Perturbations

Browse on: https://hsiung.cc/CARBEN/

Methodology

Generalized Adversarial Training (GAT)

- Objective Function



$$\min_{\theta_{\mathcal{F}}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{x_{\text{c-adv}} \in \mathcal{B}(x;\Omega;E)} \mathcal{L}(\mathcal{F}(\boxed{x_{\text{c-adv}}}), y) \right]$$

Experiments

- Baselines
  - $\text{Normal}^{\dagger}, \ \text{Normal}^{*}$: Standard training
  - $\text{Madry}^{\dagger}_{\infty}$: ℓ_{∞} adversarial training [Madry *et al.*, ICLR'18]
  - $\text{Trades}^{*}_{\infty}$: ℓ_{∞} adversarial training [Zhang *et al.*, ICML'19]
  - FAT^{*}_{∞} : adversarial training uses friendly adversarial data that are confidently misclassified [Zhang *et al.*, ICML'20]
  - AWP^{*}_{∞}: inject the worst-case weight perturbation during adversarial training to flatten the weight loss landscape [Wu *et al.*, NeurIPS'20]
  - $\text{PAT}^{\dagger}_{\text{self}}, \ \text{PAT}^{\dagger}_{\text{alex}}$: Two adversarial training models based on the perceptual distance [Laidlaw *et al.*, ICLR'21]
  - $\text{Fast-AT}^{\dagger}_{\infty}$: Computationally efficient ℓ_{∞} adversarial training by [Wong *et al.*, ICLR'21]

Experiments

- Results on CIFAR-10

Training	Clean	Three attacks CAA_{3a}	CAA_{3b}	CAA_{3c}	Semantic attacks Rand.	Sched.	Full attacks Rand.	Sched.
Normal†	95.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	59.7 ± 0.2	44.2 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
Madry$^{\dagger}_{\infty}$	87.0	30.8 ± 0.2	18.8 ± 0.5	19.1 ± 0.3	31.5 ± 0.2	21.3 ± 0.2	10.8 ± 0.2	3.7 ± 0.2
PAT$^{\dagger}_{self}$	82.4	20.9 ± 0.1	11.9 ± 0.5	17.9 ± 0.3	28.9 ± 0.3	17.5 ± 0.3	9.1 ± 0.3	2.5 ± 0.3
PAT$^{\dagger}_{alex}$	71.6	20.7 ± 0.3	12.5 ± 0.2	16.5 ± 0.4	23.4 ± 0.3	12.2 ± 0.4	10.3 ± 0.1	2.5 ± 0.2
GAT-f†	**82.3**	$\mathbf{39.9 \pm 0.1}$	$\mathbf{33.3 \pm 0.1}$	$\mathbf{28.9 \pm 0.2}$	$\mathbf{69.9 \pm 0.1}$	$\mathbf{66.0 \pm 0.1}$	$\mathbf{30.0 \pm 0.4}$	$\mathbf{18.8 \pm 0.3}$
GAT-fs†	**82.1**	$\mathbf{43.5 \pm 0.1}$	$\mathbf{36.6 \pm 0.1}$	$\mathbf{32.5 \pm 0.1}$	$\mathbf{69.9 \pm 0.2}$	$\mathbf{66.6 \pm 0.1}$	$\mathbf{32.3 \pm 0.8}$	$\mathbf{21.8 \pm 0.3}$
Normal*	94.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	46.0 ± 0.4	29.9 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
Trades$^{*}_{\infty}$	84.9	30.0 ± 0.3	19.8 ± 0.6	10.1 ± 0.5	16.6 ± 0.2	8.1 ± 0.5	5.8 ± 0.3	1.5 ± 0.2
FAT$^{*}_{\infty}$	88.1	29.8 ± 0.4	17.1 ± 0.4	12.8 ± 0.6	18.7 ± 0.2	9.8 ± 0.5	6.1 ± 0.1	1.5 ± 0.1
AWP$^{*}_{\infty}$	85.4	34.2 ± 0.2	23.2 ± 0.2	11.1 ± 0.4	15.6 ± 0.2	7.9 ± 0.2	5.9 ± 0.0	1.7 ± 0.2
GAT-f*	**83.4**	$\mathbf{40.2 \pm 0.1}$	$\mathbf{34.0 \pm 0.1}$	$\mathbf{30.7 \pm 0.4}$	$\mathbf{71.6 \pm 0.1}$	$\mathbf{67.8 \pm 0.2}$	$\mathbf{31.2 \pm 0.4}$	$\mathbf{20.1 \pm 0.3}$
GAT-fs*	**83.2**	$\mathbf{43.5 \pm 0.1}$	$\mathbf{36.3 \pm 0.1}$	$\mathbf{32.9 \pm 0.4}$	$\mathbf{70.5 \pm 0.1}$	$\mathbf{66.7 \pm 0.3}$	$\mathbf{32.2 \pm 0.7}$	$\mathbf{21.9 \pm 0.7}$

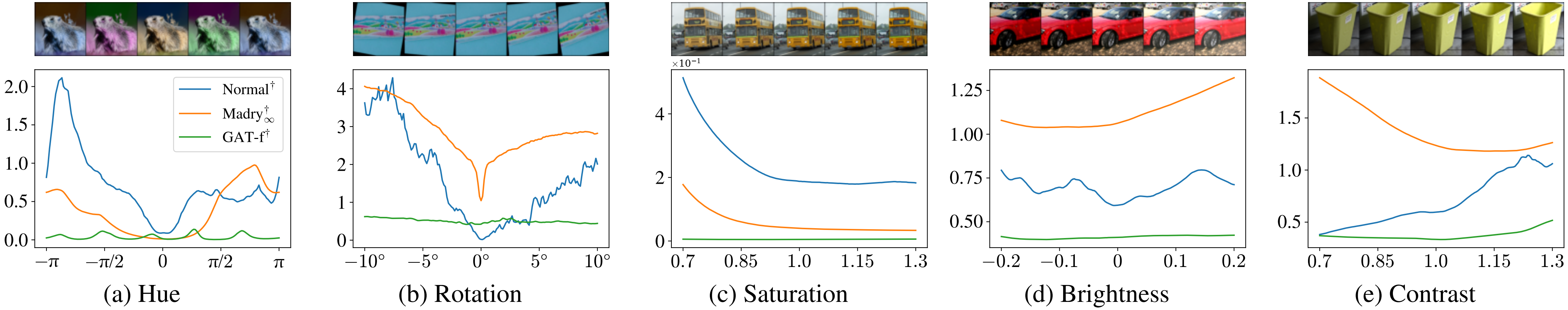
Experiments

- Results on ImageNet

Training	Clean	Three attacks CAA_{3a}	Three attacks CAA_{3b}	Three attacks CAA_{3c}	Semantic attacks Rand.	Semantic attacks Sched.	Full attacks Rand.	Full attacks Sched.
Normal†	76.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	31.2 ± 0.4	20.6 ± 1.0	0.0 ± 0.0	0.0 ± 0.0
Madry$^\dagger_\infty$	62.4	13.9 ± 0.4	9.2 ± 0.2	16.2 ± 0.8	14.0 ± 0.1	9.0 ± 0.0	7.1 ± 0.1	2.8 ± 0.2
Fast-AT$^\dagger_\infty$	53.8	9.5 ± 0.3	5.5 ± 0.1	11.4 ± 0.8	6.3 ± 0.1	3.6 ± 0.1	3.1 ± 0.1	1.0 ± 0.1
GAT-f†	60.0	**19.2 ± 1.0**	**18.9 ± 1.4**	**18.4 ± 0.4**	**43.5 ± 1.9**	**38.9 ± 2.0**	**18.5 ± 0.5**	**11.8 ± 0.1**

Discussions

- GAT's loss curve is smoother, flatter, and lower in semantic perturbation space.



Project Page

Discussions

- We maintain a leaderboard to track the robustness progress of the state-of-the-art defense method against the Composite Adversarial Attacks.

CIFAR-10

Rank	Method	Standard accuracy	AutoAttack R.A.	Semantic Attacks R.A.	Full Attacks R.A.	Architecture	Venue
1	Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations	83.2%	42.2%	66.7%	21.9%	WideResNet-34-10	CVPR 2023
2	Improving Robustness using Generated Data	85.64%	56.85%	22.21%	6.56%	WideResNet-34-20	NeurIPS 2021
3	Improving Robustness using Generated Data *It uses additional 100M synthetic images in training.*	88.74%	66.10%	17.37%	4.88%	WideResNet-70-16	NeurIPS 2021
4	Robustness and Accuracy Could Be Reconcilable by (Proper) Definition *It uses additional 1M synthetic images in training.*	87.30%	62.79%	14.83%	4.62%	ResNest-152	ICLR 2022
5	Improving Robustness using Generated Data *It uses additional 100M synthetic images in training.*	87.50%	63.44%	14.31%	4.32%	WideResNet-28-10	NeurIPS 2021
6	Fixing Data Augmentation to Improve Adversarial Robustness	88.50%	64.64%	14.36%	4.19%	WideResNet-106-16	NeurIPS 2021
7	Fixing Data Augmentation to Improve Adversarial Robustness	88.54%	64.25%	14.04%	4.11%	WideResNet-70-16	NeurIPS 2021

ImageNet

Rank	Method	Standard accuracy	AutoAttack R.A.	Semantic Attacks R.A.	Full Attacks R.A.	Architecture	Venue
1	Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations	59.96%	20.94%	38.9%	11.8%	ResNet-50	CVPR 2023
2	Towards Deep Learning Models Resistant to Adversarial Attacks *Robustness library*	62.42%	28.94%	9.0%	2.8%	ResNet-50	ICLR 2018
3	Do Adversarially Robust ImageNet Models Transfer Better?	68.41%	38.14%	9.82%	1.26%	WideResNet-50-2	NeurIPS 2020
4	Fast is better than free: Revisiting adversarial training	53.83%	24.69%	3.6%	1.0%	ResNet-50	ICLR 2020
5	Do Adversarially Robust ImageNet Models Transfer Better?	63.87%	34.96%	7.77%	0.84%	ResNet-50	NeurIPS 2020
6	Do Adversarially Robust ImageNet Models Transfer Better?	52.50%	25.32%	3.96%	0.37%	ResNet-18	NeurIPS 2020
7	Standardly trained model	76.13%	0.00%	20.6%	0.00%	ResNet-50	PyTorch

Conclusion

- With novel attack scheduling designs for multiple perturbation types, and optimizations for each attack component, CAA can easily crack modern robust models.
- CAA-generated adversarial examples enable GAT models to achieve state-of-the-art robustness against various adversarial perturbations.
- Experimental results demonstrate that GAT achieves the highest robust accuracy on most composite attacks by a large margin, providing new insights into achieving compositional adversarial robustness.
- We believe our work sheds new light on the frontiers of realistic adversarial attacks and defenses.

JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

#THU-PM-388

Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations

Lei Hsiung[1,4], Yun-Yun Tsai[2], Pin-Yu Chen[3], Tsung-Yi Ho[1,4]

[1]National Tsing Hua University [2]Columbia University [3]IBM Research [4]CUHK

國立清華大學 NATIONAL TSING HUA UNIVERSITY

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

IBM

香港中文大學 The Chinese University of Hong Kong

Project Page