Adversarially Robust Few-shot Learning via Parameter Co-distillation of Similarity and Class Concept Learners

Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, Yew-Soon Ong

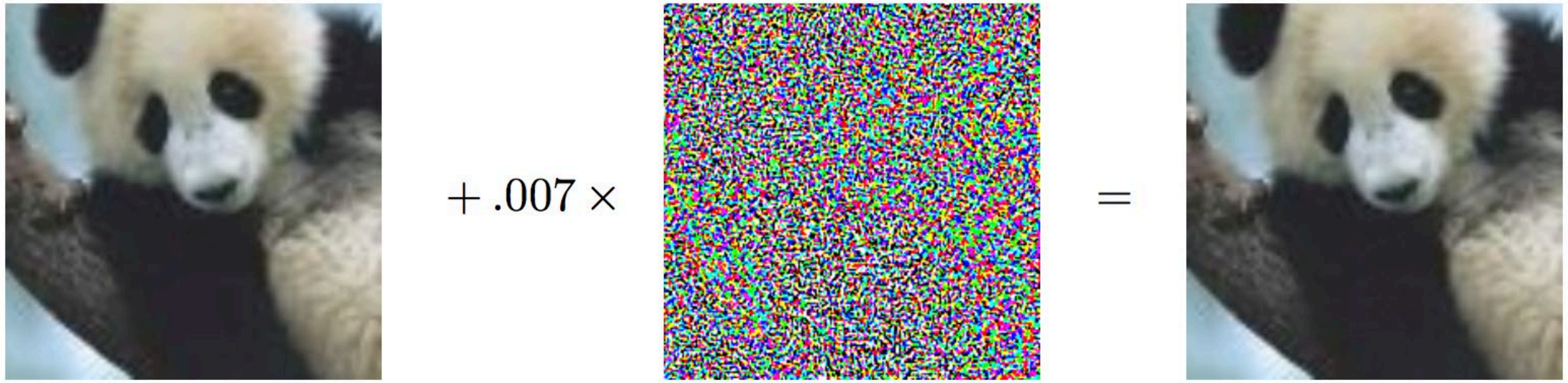
Reporter： Junhao Dong

CVPR
SEATTLE, WA
JUNE 17-21, 2024

Adversarial examples are tailored inputs with the purpose of confusing neural networks. (Visually similar to natural examples)

Adversarial Perturbation

Natural Examples



Adversarial Examples

$+ .007 \times$ $=$

$\boldsymbol{x}$
"panda"
57.7% confidence

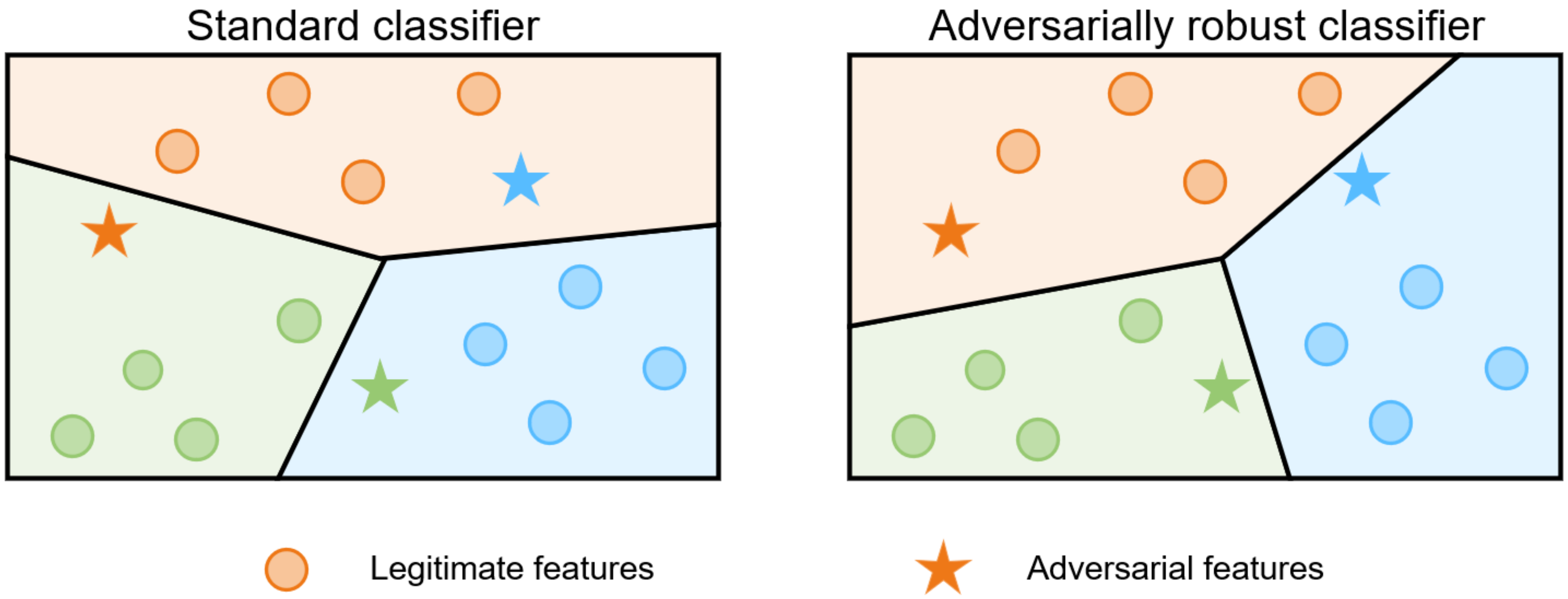
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Introducing gradient ascent at the **image level**.

Adversarial Training (min-max optimization):

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathcal{L}_{\mathrm{CE}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right),y\right)+\max_{\|\delta\|_{\infty}<\epsilon}\mathcal{L}_{\mathrm{KL}}\left(f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\|f_{\boldsymbol{\theta}}\left(\mathbf{x}+\delta\right)\right)\right]$$



Similarity learning *vs.* Class Concept Learning for Robustness

Similarity Learning:

$$\boldsymbol{\mu}_n = \frac{1}{|\mathcal{S}_n|} \sum_{(\mathbf{x},y)\in\mathcal{S}_n} f_{\boldsymbol{\theta}_s}(\mathbf{x})$$

Class-wise feature mean prototypes

$$p(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_n} | \mathbf{x}, \mathbf{M}) = \frac{\exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_n))}{\sum_{n'=1}^{N} \exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_{n'}))}$$

Learning object relations between support and query sets

Similarity learning *vs.* Class Concept Learning for Robustness

Similarity Learning:

$$\boldsymbol{\mu}_n = \frac{1}{|\mathcal{S}_n|} \sum_{(\mathbf{x},y)\in\mathcal{S}_n} f_{\boldsymbol{\theta}_s}(\mathbf{x})$$

Class-wise feature mean prototypes

$$p(y_{\mathbf{x}} = y_{\boldsymbol{\mu}_n} | \mathbf{x}, \mathbf{M}) = \frac{\exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_n))}{\sum_{n'=1}^{N} \exp(-d^2(f_{\boldsymbol{\theta}_s}(\mathbf{x}), \boldsymbol{\mu}_{n'}))}$$

Learning object relations between support and query sets

Concept Learning:

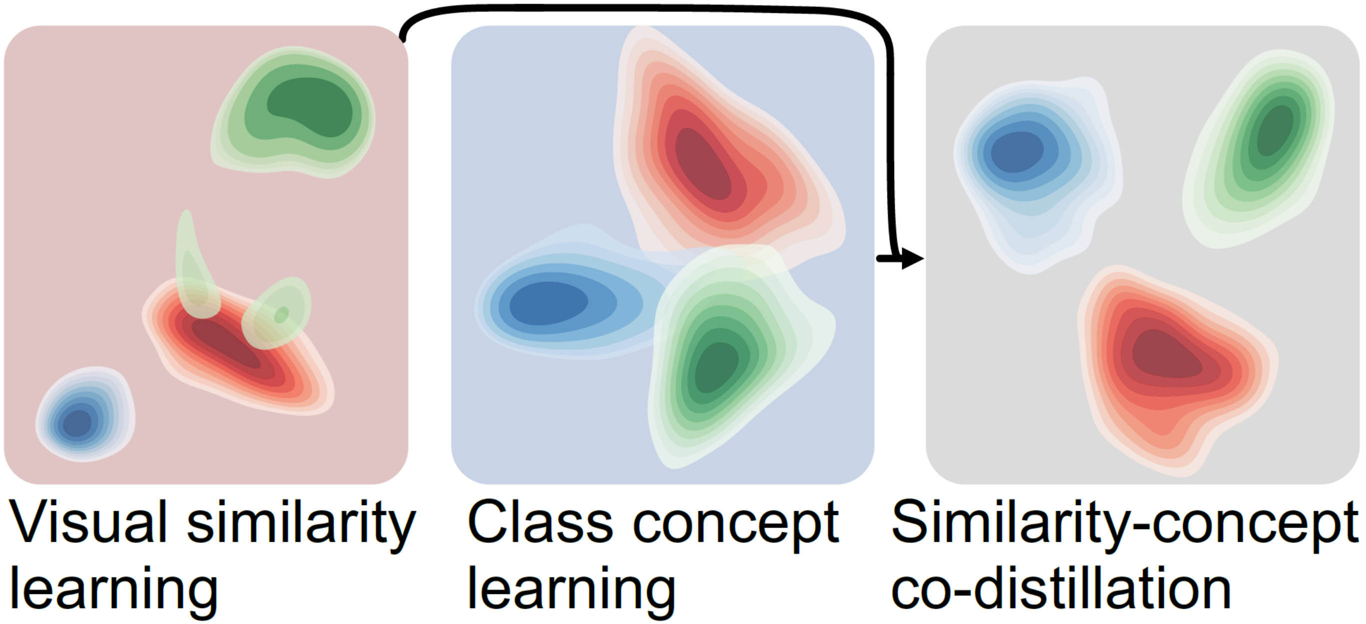
Softmax with learnable weights

$$\mathbf{W} = \{\mathbf{w}_z\}_{z=1}^{Z}$$

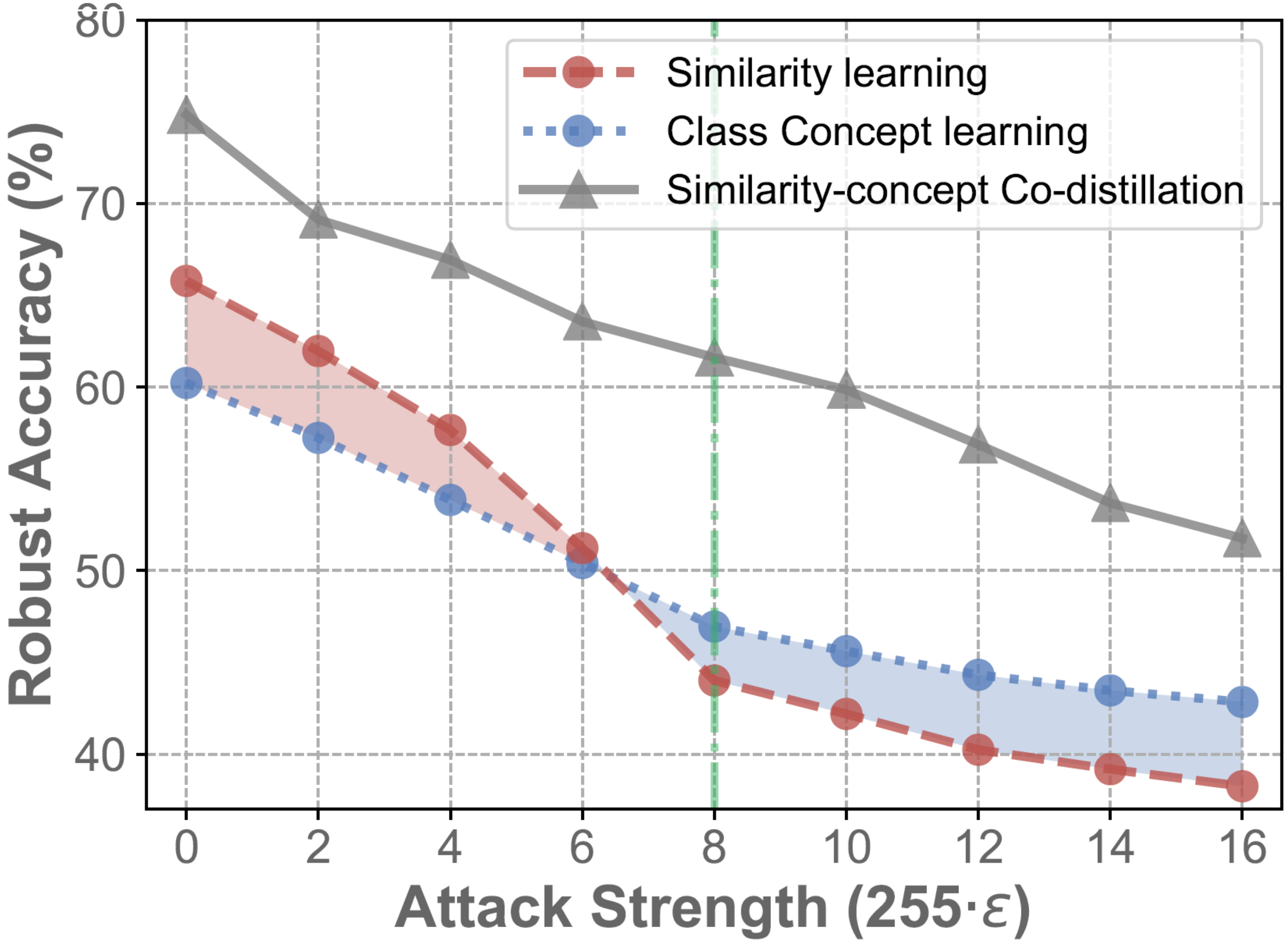
$$p(y_{\mathbf{x}} = z | \mathbf{x}, \mathbf{W}) = \frac{\exp(\mathbf{w}_z^\top f_{\boldsymbol{\theta}_c}(\mathbf{x}))}{\sum_{z'=1}^{Z} \exp(\exp(\mathbf{w}_{z'}^\top f_{\boldsymbol{\theta}_c}(\mathbf{x})))}$$

Learning global classifier weights for all the classes

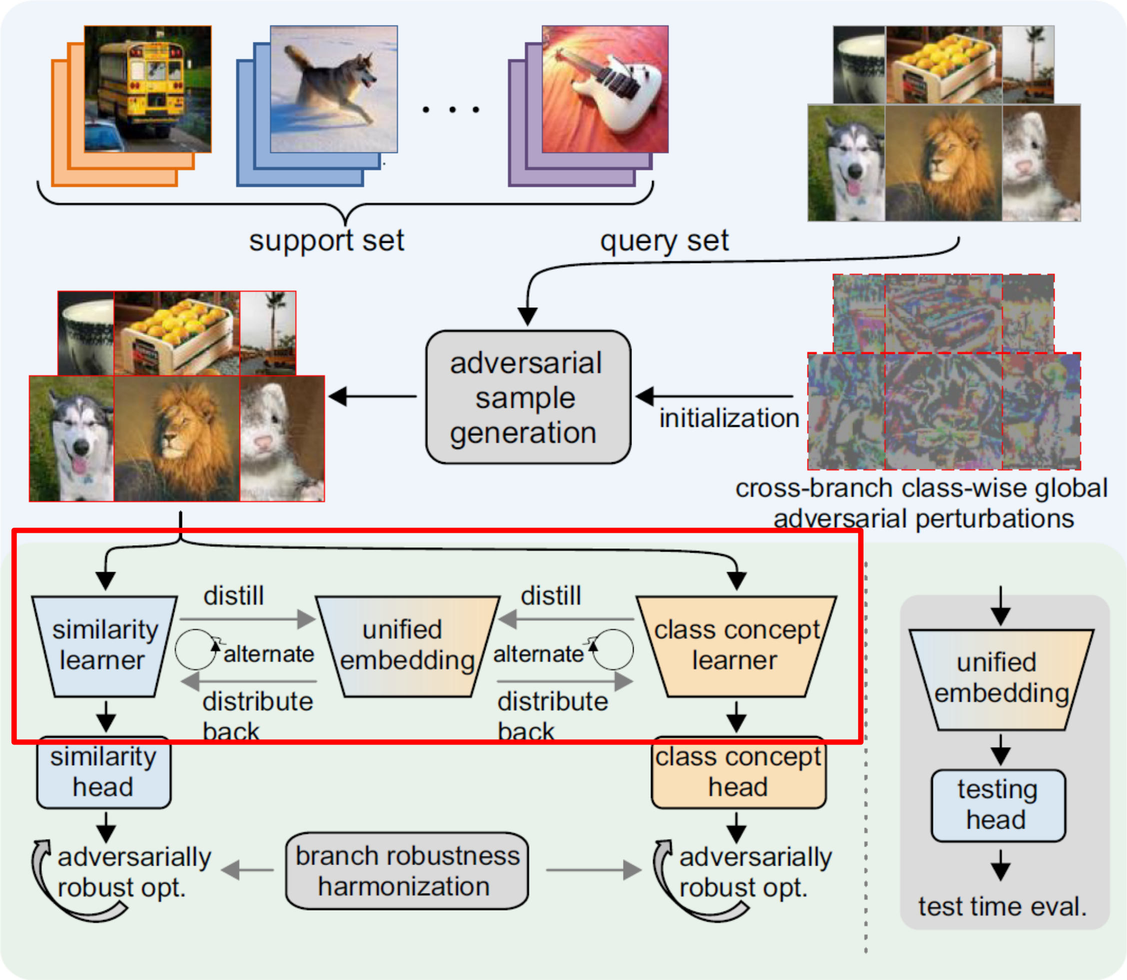
Analyses on Similarity and Class Concept Learning



■ Analyses on Similarity and Class Concept Learning



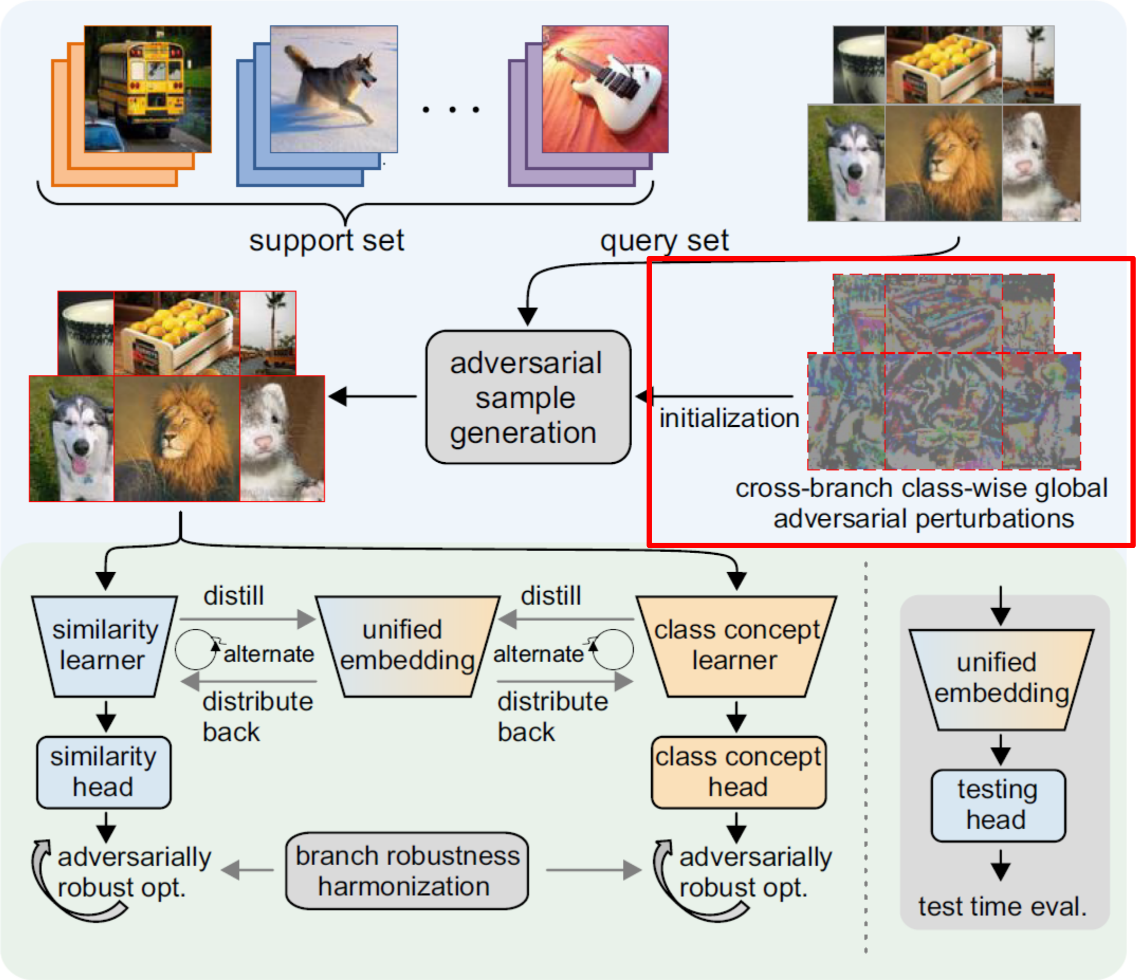
- **paRametEr co-diStIllation of SimilariTy and clAss coNCept lEarners (RESISTANCE):**



Dynamic Parameter-level Interpolation:

$$\boldsymbol{\theta}_u := \beta\boldsymbol{\theta}_u + (1-\beta)\left[\gamma\boldsymbol{\theta}_s + (1-\gamma)\boldsymbol{\theta}_c\right]$$

■ paRametEr co-diStIllation of SimilariTy and clAss coNCept lEarners (RESISTANCE):



Cross-branch Class-wise Global Adversarial Initialization Perturbations:

$$\boldsymbol{\mu}_z^{(g)} = \frac{1}{|\mathcal{B}_z|} \sum_{(\mathbf{x},y)\in\mathcal{B}_z} g(\mathbf{x})$$ Class-Wise Prototype

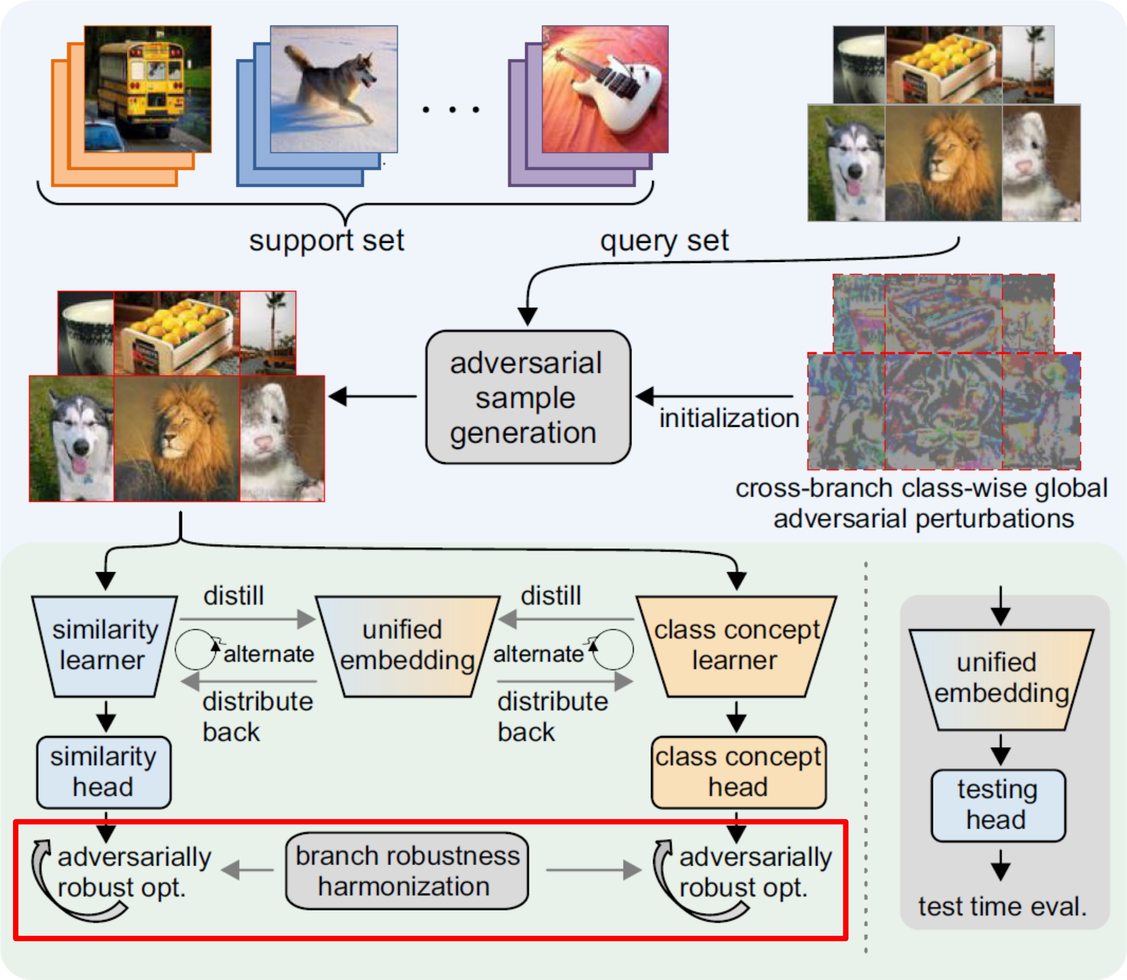
Cross-Branch Disruption

$$\mathcal{L}_{\text{GAIP}}(\mathcal{B}_z; \boldsymbol{\delta}_0^z) = \sum_{\mathbf{x}^z\in\mathcal{B}_z} \sum_{g\in\{f_{\boldsymbol{\theta}_s}, f_{\boldsymbol{\theta}_c}, f_{\boldsymbol{\theta}_u}\}} \left\| g(\mathbf{x}^z + \boldsymbol{\delta}_0^z) - \boldsymbol{\mu}_z^{(g)} \right\|_2^2$$

Iterative Perturbing

$$\boldsymbol{\delta}_0^{z\,(\iota)} = h(\mathcal{B}_z^\iota; \boldsymbol{\delta}_0^{z(\iota-1)}; \alpha) =$$
$$\Pi_{\mathbb{B}(\epsilon)}\left(\boldsymbol{\delta}_0^{z\,(\iota-1)} + \alpha\ \text{sign}\left(\nabla_{\boldsymbol{\delta}_0^{z(\iota-1)}} \mathcal{L}_{\text{GAIP}}^z(\mathcal{B}_z^\iota; \boldsymbol{\delta}_0^{z(\iota-1)})\right)\right)$$

■ **paRametEr co-diStIllation of SimilariTy and clAss coNCept lEarners (RESISTANCE):**



Branch Robustness Harmonization:

Relative Robustness Score

$$\kappa_s(\mathcal{Q},\mathcal{S}) = \frac{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\mathcal{L}_{\mathrm{KL}}\left(\mathbf{p}_{\mathbf{x}}^{\mathbf{W}} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_c^{\mathbf{x}}}^{\mathbf{W}}\right)\right]}{\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\mathcal{L}_{\mathrm{KL}}\left(\mathbf{p}_{\mathbf{x}}^{\mathbf{M}(\mathcal{S})} \| \mathbf{p}_{\mathbf{x}+\boldsymbol{\delta}_s^{\mathbf{x}}}^{\mathbf{M}(\mathcal{S})}\right)\right]}$$

Reweighted Learning Rate

$$\eta_s' = \eta_s\left[1 - \tanh\left(\tau \max\left(0, \log\left(\kappa_s\right)\right)\right)\right]$$

■ Standard Comparison:

Model	Method	Mini-ImageNet				CIFAR-FS				FC100			
		Clean	PGD	CW	AA	Clean	PGD	CW	AA	Clean	PGD	CW	AA
Conv-4	AQ [14]	50.12	28.16	27.21	24.68	57.63	39.58	38.69	37.17	35.19	24.76	22.80	21.08
	R-MAML [38]	50.76	34.19	29.61	28.31	52.75	32.66	31.47	19.25	38.56	17.67	15.91	18.75
	ST [30]	51.23	33.23	30.84	29.07	55.61	40.21	40.15	39.95	40.69	30.65	27.39	27.06
	GR [9]	50.93	37.95	35.90	31.37	58.31	47.95	46.45	45.09	41.32	32.92	30.70	29.09
	DFSL [18]	51.10	36.23	35.94	30.31	58.89	47.42	46.62	44.38	41.74	31.81	29.99	28.44
	RESISTANCE	**52.23**	**40.24**	**38.55**	**35.81**	**60.05**	**48.37**	**47.00**	**45.89**	**44.63**	**35.15**	**33.73**	**30.07**
ResNet-12	AQ [14]	64.47	30.80	29.62	25.72	65.78	44.01	42.54	41.56	41.07	25.68	24.86	22.13
	R-MAML [38]	62.75	45.78	43.88	36.12	65.61	34.77	33.15	27.77	42.25	24.39	20.49	20.08
	ST [30]	61.65	47.85	45.98	45.23	64.44	46.16	44.26	43.19	44.57	32.18	30.72	28.33
	GR [9]	64.60	50.71	47.52	47.59	66.99	52.66	50.61	50.91	46.12	34.27	32.00	30.98
	DFSL [18]	64.95	50.83	47.23	46.50	65.84	53.90	51.25	50.64	47.73	34.63	32.36	30.97
	RESISTANCE	**68.79**	**53.84**	**51.47**	**50.52**	**74.83**	**61.61**	**59.64**	**58.76**	**51.69**	**37.51**	**35.70**	**34.66**

Robustness w.r.t. diverse attack radii:

Radius ϵ	Method	Mini-ImageNet		CIFAR-FS	
		1-shot	5-shot	1-shot	5-shot
4/255	R-MAML [38]	31.67	47.21	30.96	40.43
	GR [9]	35.77	52.63	40.04	55.82
	DFSL [18]	36.39	53.45	41.12	56.92
	RESISTANCE	**39.24**	**58.57**	**46.07**	**64.18**
6/255	R-MAML [38]	28.65	42.94	27.16	34.91
	GR [9]	33.75	50.95	36.85	52.98
	DFSL [18]	33.98	50.42	37.45	53.20
	RESISTANCE	**37.06**	**54.85**	**43.39**	**61.35**
10/255	R-MAML [38]	25.08	35.73	22.81	26.12
	GR [9]	28.01	44.99	33.23	48.47
	DFSL [18]	26.83	43.08	32.98	48.03
	RESISTANCE	**29.76**	**47.33**	**38.57**	**56.18**
12/255	R-MAML [38]	23.89	32.75	21.30	25.06
	GR [9]	26.31	40.92	31.14	47.46
	DFSL [18]	25.27	38.69	29.19	45.66
	RESISTANCE	**27.65**	**44.10**	**36.01**	**52.35**

Single-step Extension (Efficiency):

Method	Adversary Type	1-shot		5-shot		Time(h)
		Clean	Robust	Clean	Robust	
R-MAML [38]	Multi-step	37.52	24.14	62.75	36.12	15.6
	N-FGSM [7]	33.61	21.27	59.72	34.53	4.8
	RS-FGSM [40]	33.86	21.22	59.85	34.48	4.8
	GradAlign [1]	34.04	21.46	60.50	34.93	8.3
GR [9]	Multi-step	45.81	32.61	64.60	47.59	10.7
	N-FGSM [7]	40.13	28.17	59.44	44.71	3.1
	RS-FGSM [40]	41.49	26.35	60.57	43.24	3.1
	GradAlign [1]	40.63	27.42	59.15	44.03	5.9
RESISTANCE	Multi-step	50.28	33.71	68.79	50.52	16.9
	N-FGSM [7]	48.84	32.70	68.40	50.35	5.3
	RS-FGSM [40]	49.24	30.26	67.81	48.70	5.3
	GradAlign [1]	49.07	31.33	68.48	49.19	9.5

Cross-domain robustness

Transfer	Method	1-shot Clean	1-shot PGD	1-shot AA	5-shot Clean	5-shot PGD	5-shot AA
M → C	AQ [14]	43.96	26.36	22.30	61.05	37.33	30.97
	GR [9]	44.13	34.67	32.13	60.86	45.17	42.03
	TROBA [17]	43.20	32.47	30.81	62.44	46.24	43.75
	RESISTANCE	**48.04**	**38.65**	**36.54**	**64.13**	**53.42**	**50.26**
M → F	AQ [14]	**36.08**	18.71	14.14	47.66	25.31	19.45
	GR [9]	35.16	26.40	24.30	45.91	33.92	30.79
	TROBA [17]	34.09	24.42	21.65	45.51	34.05	31.56
	RESISTANCE	35.78	**27.63**	**24.34**	**47.88**	**37.49**	**35.45**
C → M	AQ [14]	36.25	11.15	8.80	**56.90**	19.10	14.20
	GR [9]	36.65	24.60	20.12	50.73	33.19	30.17
	TROBA [17]	37.48	21.59	18.40	52.46	29.27	26.92
	RESISTANCE	**38.55**	**25.08**	**21.65**	56.04	**39.19**	**34.96**

Ablations:

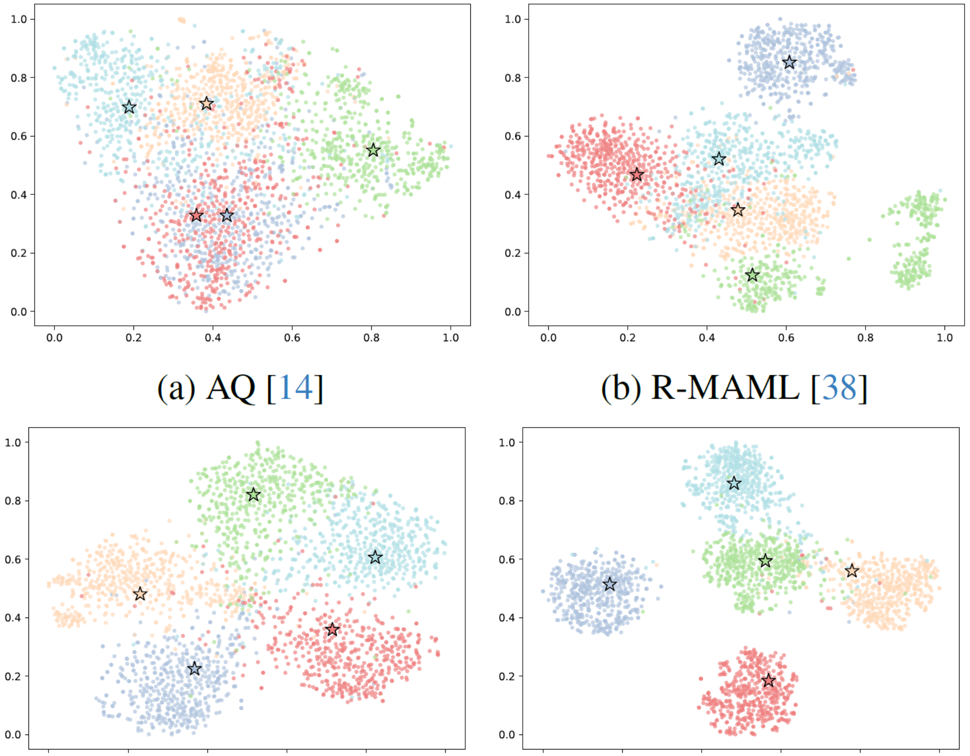
Impact of each module

	Co-dist.	GAIP	Harm.	Clean	PGD-20	AA
1				60.22	46.95	45.84
2	✓			68.12	55.14	53.07
3	✓	✓		73.17	58.99	55.72
4	✓		✓	71.46	60.24	56.20
5	✓	✓	✓	74.83	61.61	58.76

Diverse Co-distillation Components

Co-distillation Components	1-shot Clean	1-shot Robust	5-shot Clean	5-shot Robust
similarity & similarity	49.50	31.75	69.63	46.33
class concept & class concept	47.90	33.04	67.37	51.17
similarity & class concept	55.78	41.57	74.83	58.76

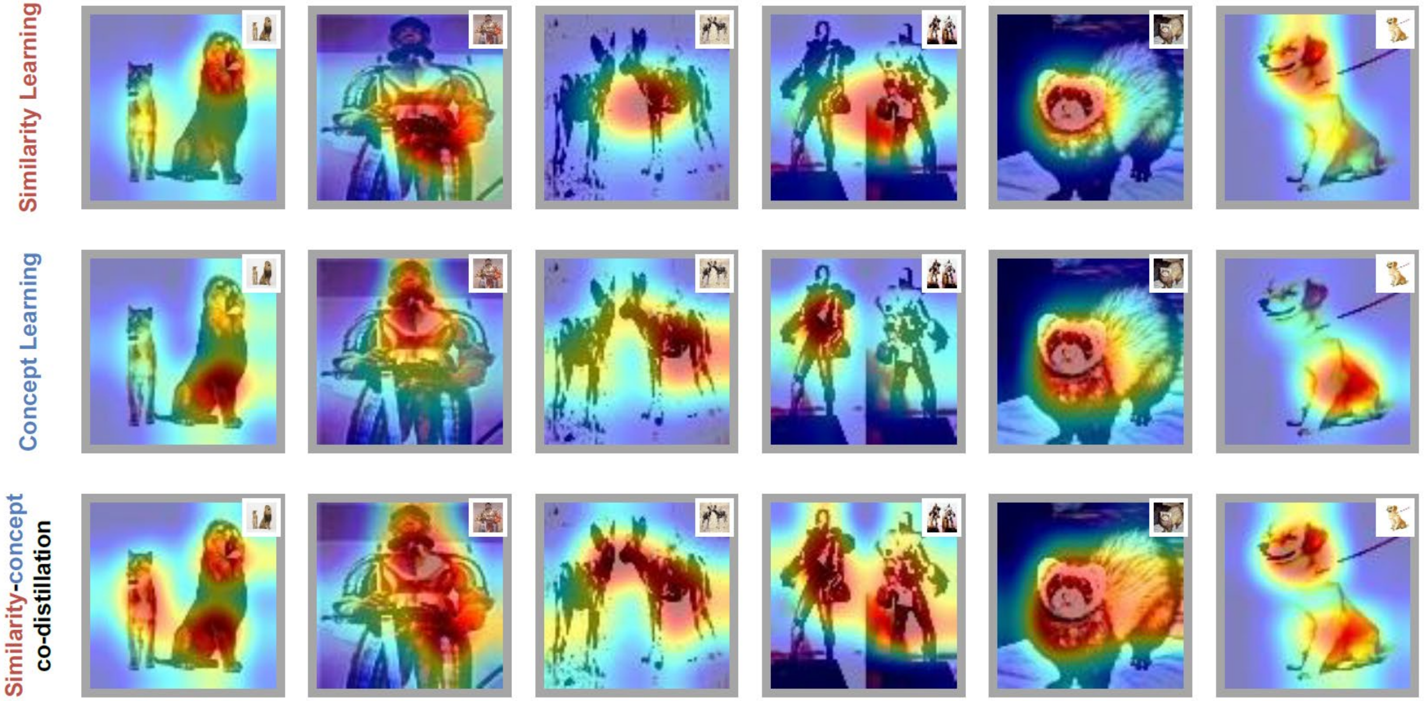
t-SNE Visualizations



(a) AQ [14] (b) R-MAML [38]

(c) GR [9] (d) **RESISTANCE**

Grad-CAM Visualizations:



■ Contributions:

- By analyzing the **complementary nature of visual similarity and class concept learning distinguished by their unique label spaces**, we propose a novel adversarially robust few-shot learning framework based on a simple but effective **parameter co-distillation** mechanism, improving robustness across diverse attack strengths.
- To promote the uniformity of robustness across learners, we introduce **cross-branch class-wise adversarial perturbations** for branch-specific adversary initialization. We also propose a **robustness harmonization** module to modulate the optimization of diverse branches.
- Comprehensive experiments demonstrate the effectiveness and generalization ability of RESISTANCE compared to the state-of-the-art adversarially robust fewshot learning approaches. In addition, we investigate the scalability of RESISTANCE with the single-step adversary generation strategies for better efficiency.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Australian National University

CSIRO

DATA 61

SEATTLE, WA

CVPR

JUNE 17-21, 2024

Thank you!

Adversarially Robust Few-shot Learning via Parameter Co-distillation of Similarity and Class Concept Learners

Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, Yew-Soon Ong

Reporter：Junhao Dong