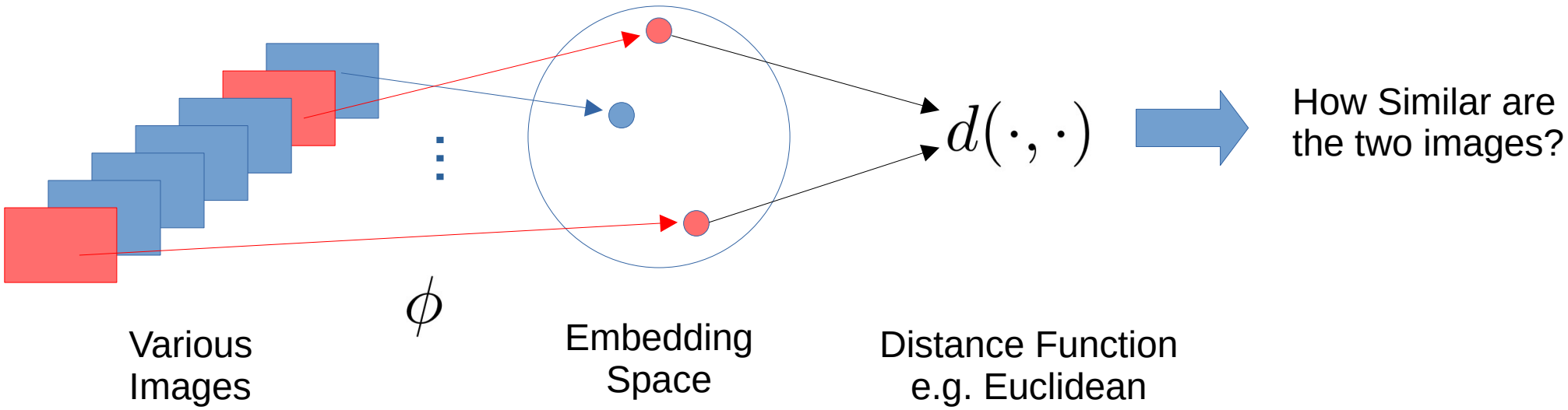
JOHNS HOPKINS
UNIVERSITY

Enhancing Adversarial Robustness for Deep Metric Learning

Mo Zhou
<mzhou32@jhu.edu>
Johns Hopkins University
Nov. 2021

Deep Metric Learning (DML)

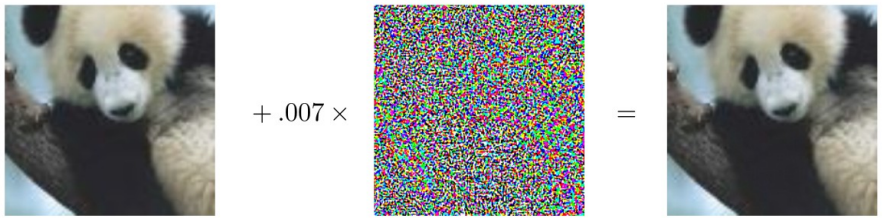
Goal: to learn a mapping function $\phi : \mathcal{X} \mapsto \Phi \subseteq \mathbb{R}^D$



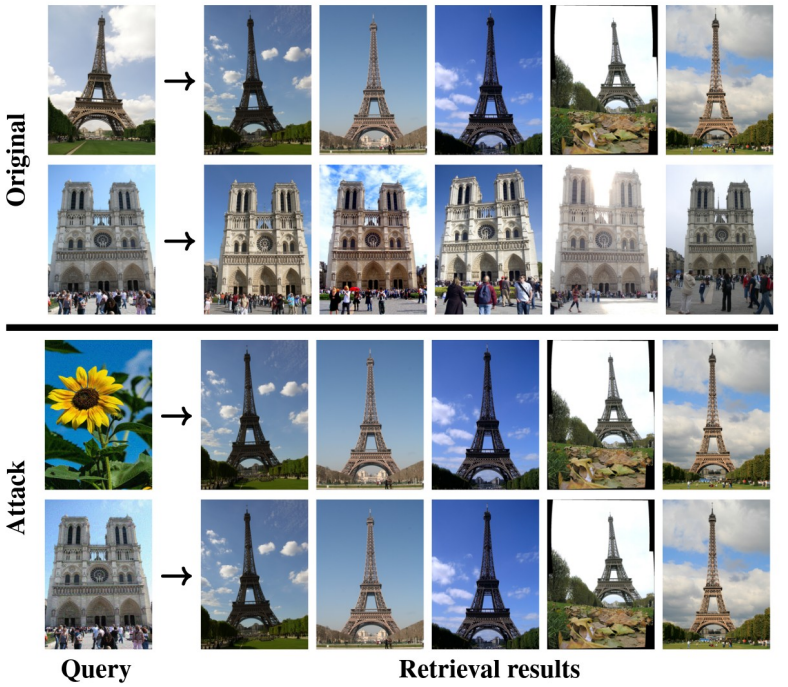
Appliations: image retrieval, face recognition, cross-modal retrieval, self-supervised learning.

[1] Karsten et al., Revisiting training strategies and generalization performance in deep metric Learning, ICML 2020.

Adversarial Attack



Attack against DNN Classifier [2]



Attack against Image Retrieval (DML) model [3]

[2] Explaining and Harnessing Adversarial Examples, ICLR 2015
[3] Giorgos et al., Targeted Mismatch Adversarial Attack: Query with a Flower to Retrieve the Tower, ICCV 2019

Adversarial Defense & Robustness for DML

- Important, but Insufficiently Explored

(1) EST Defense

Embedding-Shifted Triplet (EST) [9] adopts adversarial counterparts of $\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}$ with maximum embedding move distance off their original locations, *i.e.*,

$$L_{\text{EST}} = L_{\text{T}}(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{n}}; \gamma) \tag{1}$$

where $\tilde{\boldsymbol{a}} = \phi(\boldsymbol{A} + \boldsymbol{r}^*)$, and $\boldsymbol{r}^* = \arg\max_{\boldsymbol{r}} d_\phi(\boldsymbol{A} + \boldsymbol{r}, \boldsymbol{A})$. The $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{n}}$ are obtained similarly.

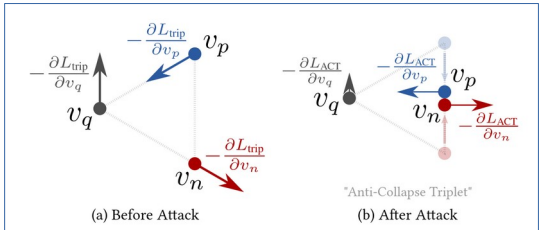


Fig. 4. Gradient Direction of Anti-Collapse Triplet (ACT). ACT does not suffer from misleading gradient or inefficient mini-batch exploitation. It does not create excessively hard adversarial example neither.

(2) ACT Defense

Anti-Collapse Triplet (ACT) [10] "collapses" the embedding vectors of positive and negative sample, and enforces the model to separate them apart, *i.e.*,

$$L_{\text{ACT}} = L_{\text{T}}(\boldsymbol{a}, \overrightarrow{\boldsymbol{p}}, \overleftarrow{\boldsymbol{n}}; \gamma), \tag{2}$$

$$[\overrightarrow{\boldsymbol{p}}, \overleftarrow{\boldsymbol{n}}] = [\phi(\boldsymbol{P} + \boldsymbol{r}_p^*), \phi(\boldsymbol{N} + \boldsymbol{r}_n^*)] \tag{3}$$

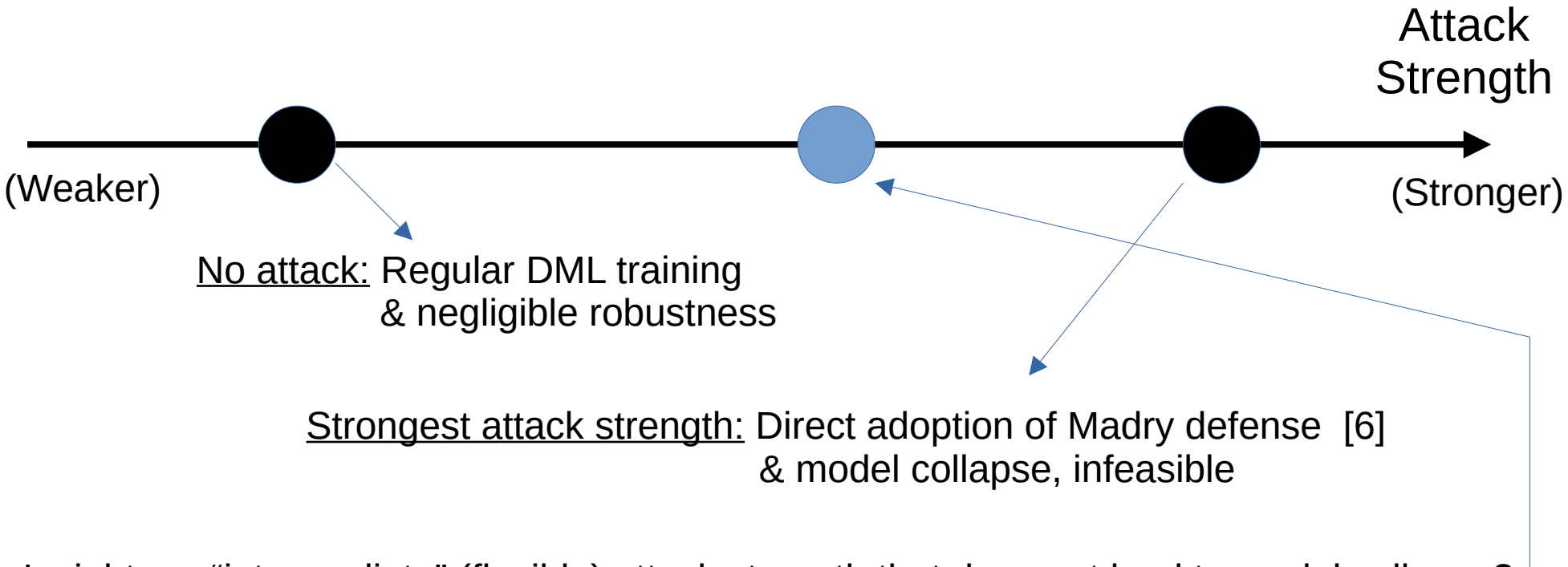
$$[\boldsymbol{r}_p^*, \boldsymbol{r}_n^*] = \arg\min_{\boldsymbol{r}_p, \boldsymbol{r}_n} d_\phi(\boldsymbol{P} + \boldsymbol{r}_p, \boldsymbol{N} + \boldsymbol{r}_n). \tag{4}$$

- **<u>Common problems: high training cost (low efficiency), performance drop, etc.</u>**

[4] M. Zhou, et al., Adversarial Ranking Attack and Defense, ECCV 2020
[5] M. Zhou, et al., Adversarial Attack and Defense in Deep Ranking, ArXiv 2021

Rethink on DML Defense -- “Interpolation”

Strengths of adversarial examples used for adversarial training as a defense:



Insight: an “intermediate” (flexible) attack strength that does not lead to model collapse?

[6] Aleksander et al., Towards deep learning models resistant to adversarial attacks, ICLR 2018

Hardness Manipulation (HM)

- "Hardness" H: Internal component of triplet loss [7]

Given an image triplet $(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{N})$ $$H(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{N}) = d_\phi(\boldsymbol{A}, \boldsymbol{P}) - d_\phi(\boldsymbol{A}, \boldsymbol{N}).$$

FYI: $$L_{\mathrm{T}}(\boldsymbol{a}, \boldsymbol{p}, \boldsymbol{n}; \gamma) = \max(0, d(\boldsymbol{a}, \boldsymbol{p}) - d(\boldsymbol{a}, \boldsymbol{n}) + \gamma),$$ (triplet loss)

- Hardness Manipulation: flexible tool for creating adversarial examples

(1) hardness of adversarial triplet $$\tilde{H}_{\mathsf{S}} = H(\boldsymbol{A}+\boldsymbol{r}_a, \boldsymbol{P}+\boldsymbol{r}_p, \boldsymbol{N}+\boldsymbol{r}_n)$$

(2) projected gradient descent (create adv examples) $$\hat{\boldsymbol{r}}_a, \hat{\boldsymbol{r}}_p, \hat{\boldsymbol{r}}_n = \underset{\boldsymbol{r}_a, \boldsymbol{r}_p, \boldsymbol{r}_n}{\arg\min} \left\| \max(0, H_{\mathsf{D}} - \tilde{H}_{\mathsf{S}}) \right\|_2^2.$$

(3) optimize model parameters (adversarial training) $$L_{\mathrm{T}}(\phi(\boldsymbol{A} + \hat{\boldsymbol{r}}_a), \phi(\boldsymbol{P} + \hat{\boldsymbol{r}}_p), \phi(\boldsymbol{N} + \hat{\boldsymbol{r}}_n))$$

[7] Florian et al., Facenet: A unified embedding for face recognition and clustering. CVPR 2015

HM: Characteristics

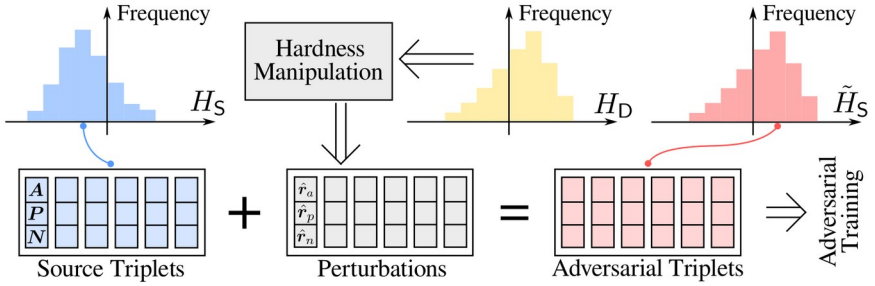
- Direct & Efficient in reaching the "intermediate" point.

$$\Delta \boldsymbol{r} = \operatorname{sign}\Big\{ -\frac{\partial}{\partial \boldsymbol{r}} \big\| \max(0, H_{\mathsf{D}} - \tilde{H}_{\mathsf{S}}) \big\|_2^2 \Big\} \quad (4)$$

$$= \operatorname{sign}\Big\{ 2(H_{\mathsf{D}} - \tilde{H}_{\mathsf{S}}) \frac{\partial}{\partial \boldsymbol{r}} \tilde{H}_{\mathsf{S}} \Big\} = \operatorname{sign}\Big\{ \frac{\partial}{\partial \boldsymbol{r}} \tilde{H}_{\mathsf{S}} \Big\}. \quad (5)$$

$$\boldsymbol{r} \leftarrow \operatorname{Proj}_{\Gamma}\{\boldsymbol{r} + \alpha \Delta \boldsymbol{r}\}$$

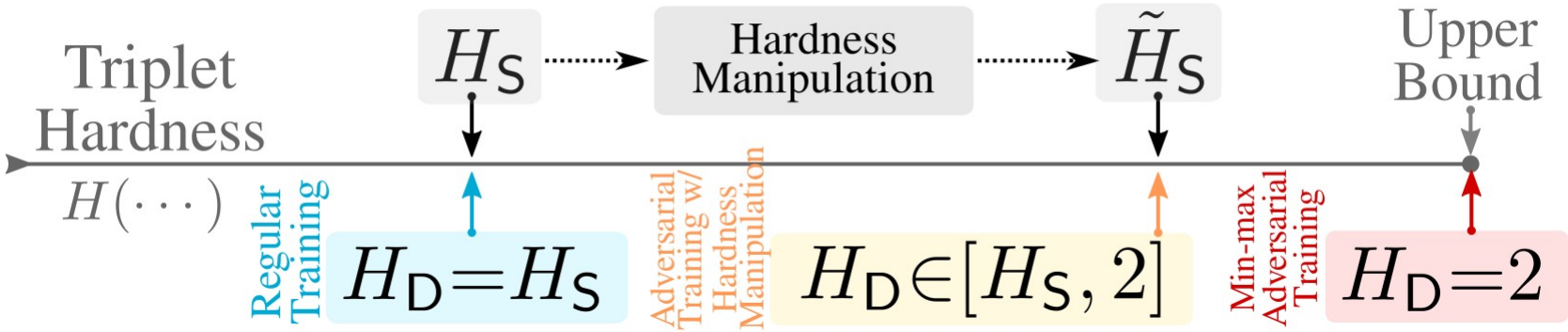
The same sign of gradient as that for directly maximizing the hardness (Madry case).



Interpretation: direct maximization process of the hardness which stops "halfway" at the specified destination hardness, instead of pushing it towards theoretical upper bound.

HM: Destination Hardness H_{D}

- Constant (not flexible)
- Hardness of another benign triplet (unbalanced)
- Pseudo-hardness function (requires prior-knowledge)



Experiment: Selection of Hardness

- Start from Hardness of another benign triplet. Because these harder but benign triplets does not lead to model collapse in regular DML training.

(1) sorting triplet sampling strategies based on mean & variance

Statistics	Random	Semihard	Softhard	Distance	Hardest
$E[H]$	-0.164	-0.126	-0.085	0.043	0.044
$\mathrm{Var}[H]$	0.00035	0.00013	0.00122	0.00021	0.00021

Table 1. Mean & variance of hardness w/ various triplet samplers. Calculated as the average statistics over 1000 mini-batches from the CUB dataset with an imagenet-initialized RN18 model.

(2) experiment all combinations

H_S \ H_D	Random		Semihard		Softhard		Distance		Hardest	
	R@1	ERS	R@1	ERS	R@1	ERS	R@1	ERS	R@1	ERS
Random	53.9	3.8	27.0	35.1	Collapse		Collapse		Collapse	
Semihard	43.9	5.4	44.0	5.0	Collapse		Collapse		Collapse	
Softhard	48.3	13.7	38.4	29.6	55.7	6.2	Collapse		Collapse	
Distance	52.7	4.8	50.7	4.8	Collapse		51.4	4.9	54.7	5.4
Hardest	51.0	4.7	52.2	4.8	Collapse		52.6	5.1	48.9	5.0

Table 2. Combinations of source & destination hardness. Evaluated on the CUB Dataset with RN18 model. The last-epoch performance is reported instead of the peak performance for alignment. Models on the diagonal are regularly (instead of adversarially) trained.

Experiment: Two cases with good performance

Dataset	Defense	η	Benign Example				White-Box Attacks for Robustness Evaluation										ERS↑
			R@1↑	R@2↑	mAP↑	NMI↑	CA+↑	CA-↓	QA+↑	QA-↓	TMA↓	ES:D↓	ES:R↑	LTM↑	GTM↑	GTT↑	
CUB	N/A[$\mathcal{R}$]	N/A	53.9	66.4	26.1	59.5	0.0	100.0	0.0	99.9	0.883	1.762	0.0	0.0	14.1	0.0	3.8
CUB	ACT[$\mathcal{R}$] [53]	2	46.5	58.4	29.1	55.6	0.6	98.9	0.4	98.1	0.837	1.666	0.2	0.2	19.6	0.0	5.8
	ACT[$\mathcal{R}$] [53]	4	38.4	49.8	22.8	49.7	4.6	81.9	2.8	80.5	0.695	1.366	2.9	2.3	18.8	0.1	13.9
	ACT[$\mathcal{R}$] [53]	8	30.6	40.1	16.5	45.6	13.7	46.8	12.6	39.3	0.547	0.902	13.6	9.8	21.9	1.3	31.3
	ACT[$\mathcal{R}$] [53]	16	28.6	38.7	15.1	43.7	15.8	37.9	16.0	31.5	0.496	0.834	11.3	9.8	21.2	2.1	34.7
	ACT[$\mathcal{R}$] [53]	32	27.5	38.2	12.2	43.0	15.5	37.7	15.1	32.2	0.472	0.821	11.1	9.4	14.9	1.0	33.9
CUB	ACT[$\mathcal{S}$] [53]	2	53.0	65.1	34.7	59.9	0.0	100.0	0.0	99.8	0.877	1.637	0.0	0.0	20.4	0.0	5.1
	ACT[$\mathcal{S}$] [53]	4	49.3	61.0	31.5	56.6	0.6	97.6	0.2	98.1	0.799	1.485	0.3	0.2	18.9	0.0	7.1
	ACT[$\mathcal{S}$] [53]	8	42.8	54.7	26.6	53.3	4.8	72.8	2.7	73.3	0.619	1.148	8.3	4.9	23.5	0.3	18.7
	ACT[$\mathcal{S}$] [53]	16	40.5	51.6	24.8	51.7	6.7	62.1	4.9	60.6	0.566	1.014	12.4	8.6	22.5	0.9	23.7
	ACT[$\mathcal{S}$] [53]	32	39.4	50.2	18.6	51.3	6.8	61.5	5.2	60.4	0.506	1.032	12.8	11.3	17.7	0.3	24.2
CUB	HM[$\mathcal{R}, \mathcal{M}$]	2	34.3	44.9	19.5	47.4	7.7	77.5	6.5	70.8	0.636	1.281	4.3	2.6	21.1	0.2	18.1
	HM[$\mathcal{R}, \mathcal{M}$]	4	30.7	40.3	16.4	45.3	13.9	60.4	13.5	48.1	0.582	1.041	6.6	6.6	20.2	1.2	27.1
	HM[$\mathcal{R}, \mathcal{M}$]	8	27.0	36.0	13.2	42.5	19.4	48.0	22.2	32.0	0.535	0.867	11.6	10.4	19.3	2.9	35.1
	HM[$\mathcal{R}, \mathcal{M}$]	16	23.8	32.6	11.6	40.6	20.9	45.0	24.6	28.6	0.494	0.805	15.6	11.3	22.1	3.2	38.0
	HM[$\mathcal{R}, \mathcal{M}$]	32	23.1	31.9	11.3	40.3	22.8	46.0	24.3	28.3	0.495	0.800	14.2	11.7	19.7	3.8	38.0
CUB	HM[$\mathcal{S}, \mathcal{M}$]	2	44.5	56.1	27.8	53.3	1.9	87.7	1.6	88.8	0.827	1.101	3.7	0.3	19.0	0.0	11.6
	HM[$\mathcal{S}, \mathcal{M}$]	4	40.6	51.8	24.2	51.0	7.3	64.1	6.3	60.9	0.715	0.894	7.9	4.4	22.8	0.2	22.1
	HM[$\mathcal{S}, \mathcal{M}$]	8	38.4	49.7	22.9	50.3	10.9	50.5	10.8	44.6	0.680	0.722	13.3	11.2	25.8	1.2	29.6
	HM[$\mathcal{S}, \mathcal{M}$]	16	37.4	47.3	21.0	48.2	14.4	42.0	14.8	34.7	0.599	0.693	17.5	14.4	26.5	2.4	34.8
	HM[$\mathcal{S}, \mathcal{M}$]	32	35.3	46.1	20.2	48.0	15.1	41.8	15.2	33.0	0.589	0.686	18.7	14.9	27.8	2.9	35.7

Table 3. Hardness manipulation in adversarial training. The "↑" mark means "the higher the better", while "↓" means the opposite.

Experiment: Curves for the previous page

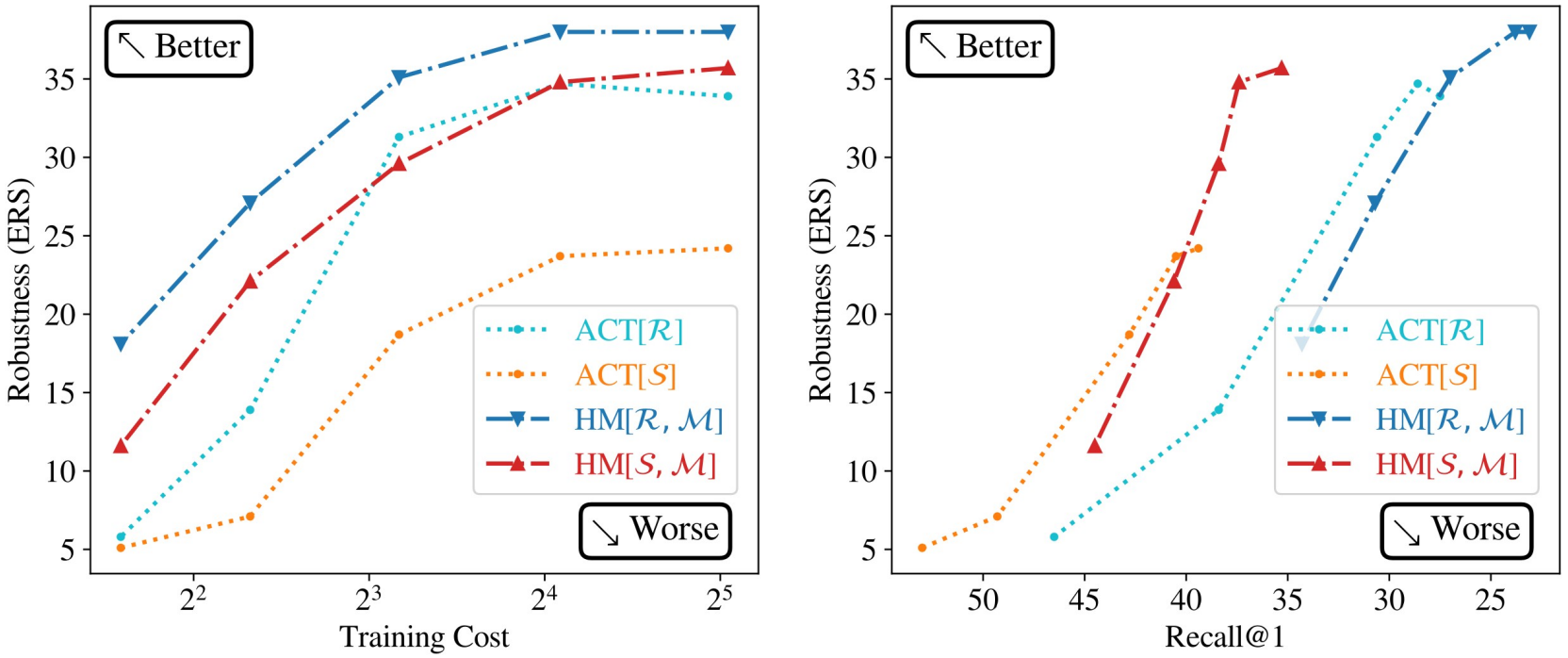


Figure 6. Performance of HM[$\mathcal{R}$, $\mathcal{M}$] & HM[$\mathcal{S}$, $\mathcal{M}$] in Tab. 3.

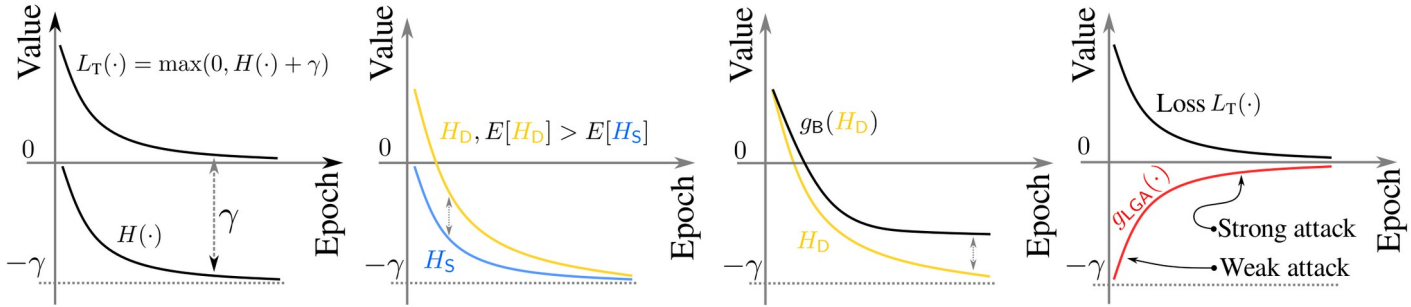
Gradual Adversary

- Existing issues

  (1) adv example insufficiently strong in the late phase of training (shown by subfigure (b) below)

  → can be alleviated by slightly boosting H_D in the late phase of training (shown by subfigure(c))

  (2) adv example being too strong in the early phase of training when the embedding space is not “in-shape”.



(a) Relationship between loss and hardness when $L_T(\cdot)>0$

(b) Shrinked $E[H_D - H_S]$ value leads to weak attack

(c) Slightly Boosting H_D for slightly stronger attack

(d) Linear Graduate Adversary for balanced training

- We propose Linear Gradual Adversary, a pseudo-hardness function for HM

$$g_{\mathsf{LGA}}(\bar{\ell}_{t-1}) = -\gamma \cdot \bar{\ell}_{t-1} \in [-\gamma, 0].$$

Dataset	Defense	η	Benign Example				White-Box Attacks for Robustness Evaluation									ERS↑
			R@1↑	R@2↑	mAP↑	NMI↑	CA+↑	CA-↓	QA+↑	QA-↓	TMA↓	ES:D↓	ES:R↑	LTM↑	GTM↑	GTT↑	
CUB	HM[S, $\mathcal{M}$]	8	38.4	49.7	22.9	50.3	10.9	50.5	10.8	44.6	0.680	0.722	13.3	11.2	25.8	1.2	29.6
	HM[S, $g_{\mathrm{B}}(\mathcal{M})$] ($\xi = 0.1$)	8	36.5	48.0	21.4	48.4	13.0	44.0	13.2	35.6	0.667	0.628	20.3	13.2	26.7	2.8	33.8
	HM[S, 0]	8	0.8	0.9	0.8	6.0	19.8	92.4	42.0	51.9	1.000	0.000	1.2	1.2	1.0	14.1	29.7
	HM[S, $-\gamma/2$]	8	36.8	47.9	21.7	48.5	12.7	41.5	12.2	35.7	0.668	0.633	18.1	14.3	28.4	2.9	33.8
	HM[S, $-\gamma$]	8	37.8	48.1	22.1	48.7	11.7	48.4	11.3	43.2	0.541	0.850	15.2	11.6	26.1	1.3	31.2
	HM[S, g_{LGA}]	8	38.0	48.3	21.8	49.3	12.7	46.4	11.6	39.9	0.567	0.783	16.8	11.9	27.9	1.4	32.4
CUB	HM[S, $-\gamma/2$]	2	44.5	55.9	27.6	53.3	2.4	86.1	1.3	87.7	0.809	1.091	1.5	1.8	22.1	0.1	12.4
	HM[S, $-\gamma/2$]	4	40.0	50.7	23.8	50.3	8.0	59.8	7.5	55.5	0.694	0.860	10.2	6.5	26.2	0.4	24.7
	HM[S, $-\gamma/2$]	8	36.8	47.9	21.7	48.5	12.7	41.5	12.2	35.7	0.668	0.633	18.1	14.3	28.4	2.9	33.8
	HM[S, $-\gamma/2$]	16	35.2	45.8	20.3	47.6	15.0	36.3	15.1	30.7	0.638	0.595	19.6	16.8	28.7	3.5	36.8
	HM[S, $-\gamma/2$]	32	34.7	45.5	20.0	47.5	15.0	36.7	15.1	29.9	0.631	0.611	20.1	17.2	29.3	3.5	37.0
CUB	HM[S, g_{LGA}]	2	47.5	59.3	30.1	55.3	1.8	88.1	1.1	88.9	0.854	1.022	2.3	0.8	21.2	0.0	11.7
	HM[S, g_{LGA}]	4	42.7	53.6	26.3	52.6	6.5	67.3	4.6	65.0	0.734	0.893	6.6	5.8	23.7	0.3	20.8
	HM[S, g_{LGA}]	8	38.0	48.3	21.8	49.3	12.7	46.4	11.6	39.9	0.567	0.783	16.8	11.9	27.9	1.4	32.4
	HM[S, g_{LGA}]	16	37.0	47.2	21.3	48.4	13.6	42.2	13.1	35.9	0.533	0.757	16.3	15.3	27.2	2.1	34.5
	HM[S, g_{LGA}]	32	36.5	46.7	21.0	48.6	14.7	39.6	15.6	34.2	0.523	0.736	16.5	15.0	26.7	2.9	35.9

Experiment: Linear Gradual Adversary

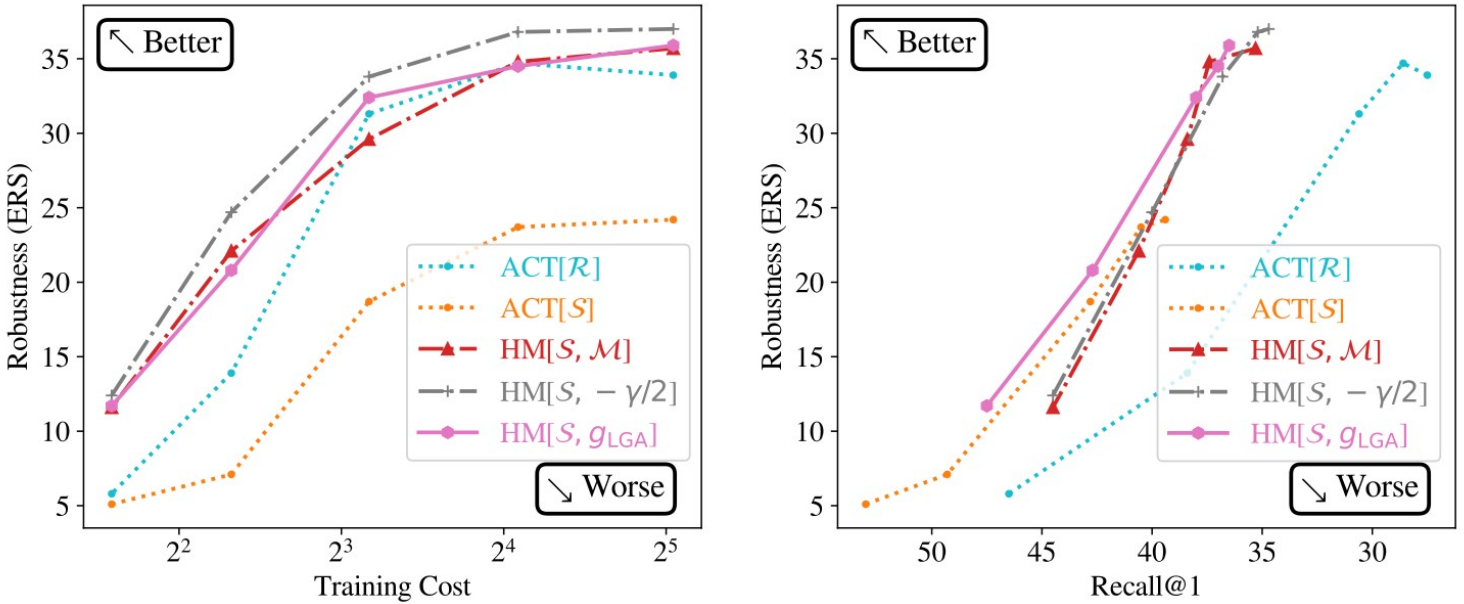
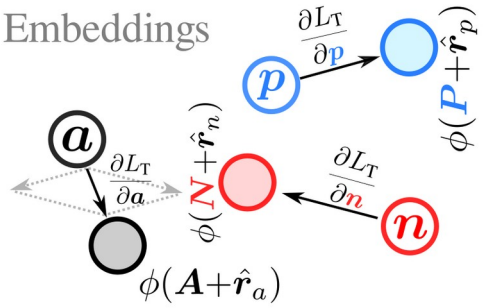


Figure 7. Performance of "HM[$\mathcal{S}$, g_{LGA}]" in Tab. 4.

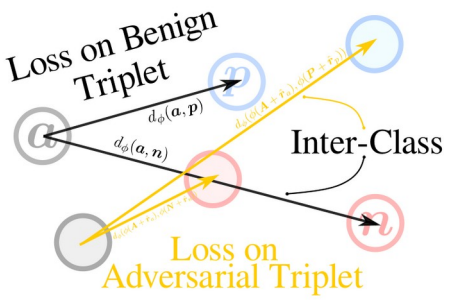
Nobody really reads such bulky table from slides.
It is provided just as a reference.

Intra-Class Structure (ICS) Loss

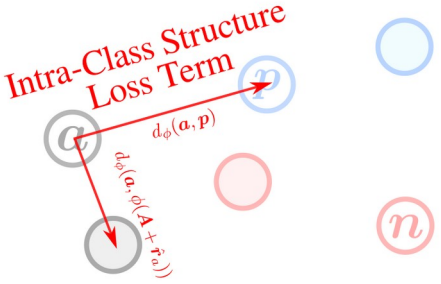
- Benign triplet + Adversarial triplet = Sextuplet (i.e., 6-item tuple).
- Enforcing intra-class structure is possible
- There are attacks aiming to change item ranking within the same class
- Neglected by previous defense methods.



(a) Benign & Adversarial Triplets = Sextuplet



(b) Loss on Benign and Adversarial Triplets



(c) Intra-Class Structure Loss Term L_{ICS}

- ICS loss term: $L_{\text{ICS}} = \lambda \cdot L_{\text{T}}(\boldsymbol{a}, \phi(\boldsymbol{A} + \hat{\boldsymbol{r}}_a), \boldsymbol{p}; 0),$

Dataset	Defense	η	Benign Example				White-Box Attacks for Robustness Evaluation								ERS↑
			R@1↑	R@2↑	mAP↑	NMI↑	CA+↑	CA-↓	QA+↑	QA-↓	TMA↓	ES:D↓	ES:R↑	LTM↑	GTM↑	GTT↑	
CUB	HM[$\mathcal{R}$, $\mathcal{M}$]	8	27.0	36.0	13.2	42.5	19.4	48.0	22.2	32.0	0.535	0.867	11.6	10.4	19.3	2.9	35.1
	HM[$\mathcal{R}$, $\mathcal{M}$]&ICS	8	25.6	34.3	12.5	41.8	21.9	41.0	23.6	26.4	0.497	0.766	14.5	13.0	21.8	4.7	39.0
CUB	HM[$\mathcal{S}$, $\mathcal{M}$]	8	38.4	49.7	22.9	50.3	10.9	50.5	10.8	44.6	0.680	0.722	13.3	11.2	25.8	1.2	29.6
	HM[$\mathcal{S}$, $\mathcal{M}$]&ICS	8	36.9	48.9	21.6	48.8	12.4	42.9	12.5	36.6	0.850	0.446	17.0	13.9	27.2	1.9	32.3
CUB	HM[$\mathcal{R}$, g_{LGA}]	8	24.8	33.9	12.2	41.6	21.4	45.0	21.7	31.3	0.452	0.846	13.2	12.0	20.9	4.6	37.3
	HM[$\mathcal{R}$, g_{LGA}]&ICS	8	25.7	35.2	12.8	41.7	22.1	37.1	23.4	23.7	0.464	0.725	14.5	13.3	21.1	5.3	40.2
CUB	HM[$\mathcal{S}$, g_{LGA}]	8	38.0	48.3	21.8	49.3	12.7	46.4	11.6	39.9	0.567	0.783	16.8	11.9	27.9	1.4	32.4
	HM[$\mathcal{S}$, g_{LGA}]&ICS	8	37.2	47.8	21.4	48.4	12.9	40.9	14.7	33.7	0.806	0.487	17.1	13.2	26.3	2.3	33.5
	HM[$\mathcal{S}$, g_{LGA}]&ICS(λ=1.0)	8	36.0	46.7	20.7	48.0	14.2	41.0	15.1	31.7	0.907	0.329	17.0	14.2	24.5	2.1	33.7
CUB	HM[$\mathcal{S}$, g_{LGA}]&ICS	2	45.2	57.2	28.5	53.7	3.0	79.9	2.4	78.9	0.936	0.609	3.6	1.2	19.9	0.0	15.2
	HM[$\mathcal{S}$, g_{LGA}]&ICS	4	41.8	53.0	25.3	52.0	8.1	57.3	7.9	54.1	0.892	0.514	9.8	6.7	22.9	0.5	24.6
	HM[$\mathcal{S}$, g_{LGA}]&ICS	8	37.2	47.8	21.4	48.4	12.9	40.9	14.7	33.7	0.806	0.487	17.1	13.2	26.3	2.3	33.5
	HM[$\mathcal{S}$, g_{LGA}]&ICS	16	35.5	46.4	20.4	47.5	14.9	37.2	17.1	30.3	0.771	0.495	18.2	15.3	28.7	2.8	36.0
	HM[$\mathcal{S}$, g_{LGA}]&ICS	32	34.9	45.0	19.8	47.1	15.5	37.7	16.6	30.9	0.753	0.506	17.9	16.7	27.3	2.9	36.0

Experiment: ICS Loss

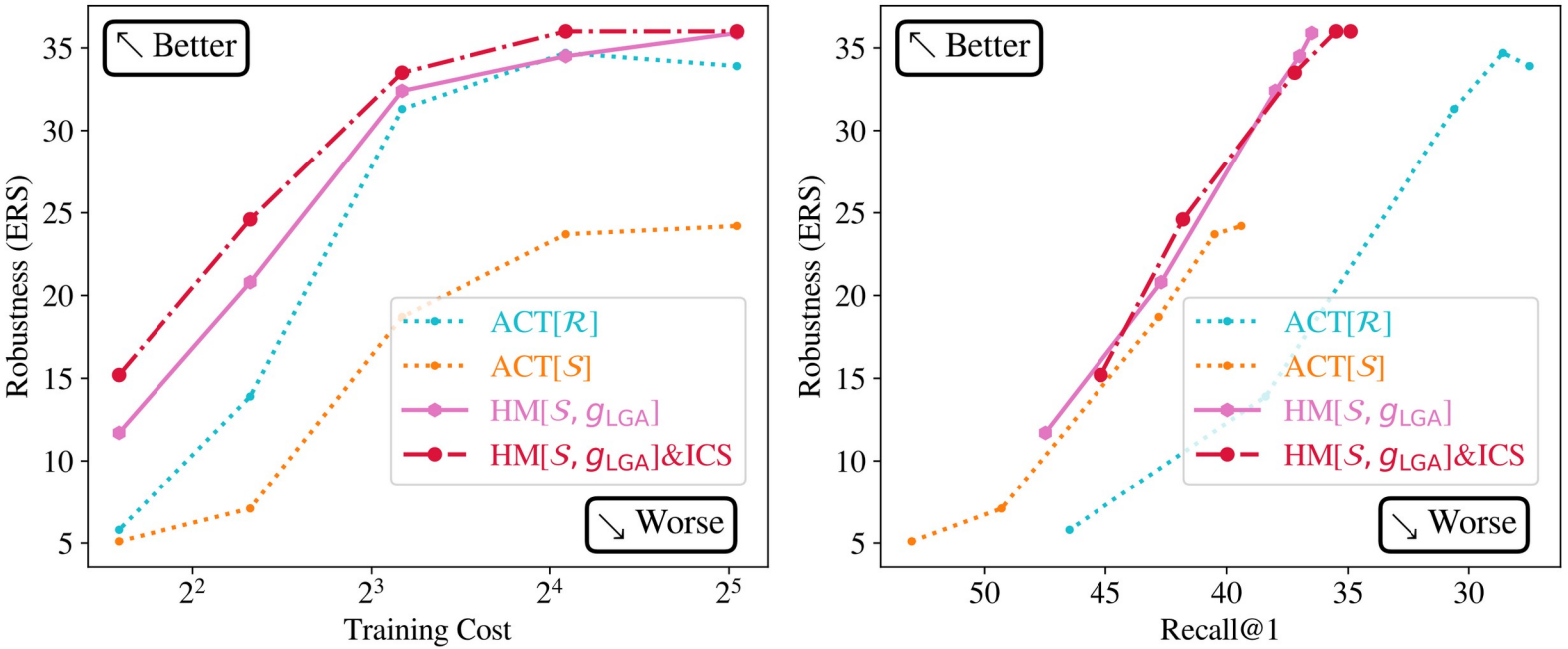


Figure 8. Performance of "HM[$\mathcal{S}$, g_{LGA}]&ICS" in Tab. 5.

Simplified Plot: all improvements combined

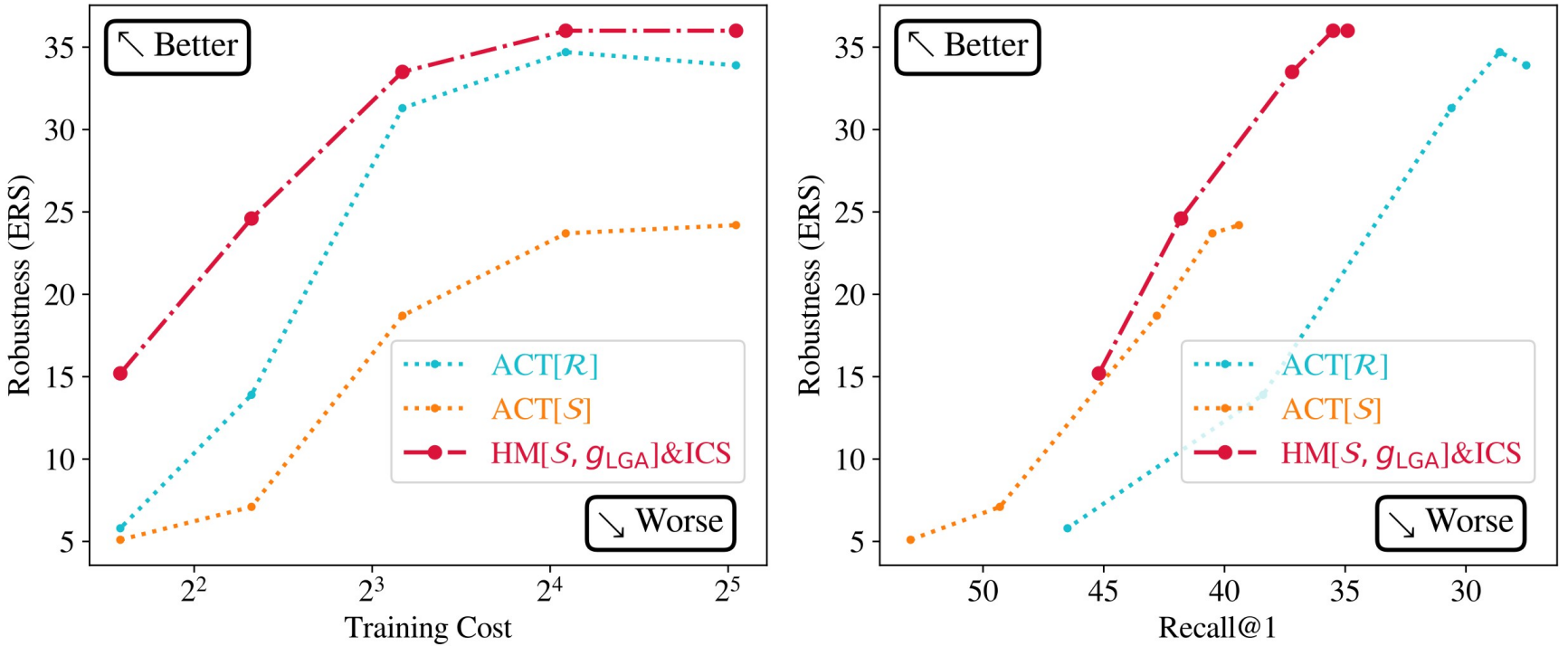


Figure 1. Comparison in robustness, training cost, and recall@1 between our method (*i.e.*, "HM[$\mathcal{S}$, g_{LGA}]&ICS") and the state-of-the-art method (*i.e.*, "ACT[$\mathcal{R}$]" and "ACT[$\mathcal{S}$]") on the CUB Dataset.

Comparison with SoTA

Dataset	Defense	η	Benign Example				White-Box Attacks for Robustness Evaluation										ERS↑
			R@1↑	R@2↑	mAP↑	NMI↑	CA+↑	CA-↓	QA+↑	QA-↓	TMA↓	ES:D↓	ES:R↑	LTM↑	GTM↑	GTT↑	
CUB	N/A[$\mathcal{R}$]	N/A	53.9	66.4	26.1	59.5	0.0	100.0	0.0	99.9	0.883	1.762	0.0	0.0	14.1	0.0	3.8
	EST[$\mathcal{R}$] [51]	8	37.1	47.3	20.0	46.4	0.5	97.3	0.5	91.3	0.875	1.325	3.9	0.4	14.9	0.0	7.9
	ACT[$\mathcal{R}$] [53]	8	30.6	40.1	16.5	45.6	13.7	46.8	12.6	39.3	0.547	0.902	13.6	9.8	21.9	1.3	31.3
	HM[$\mathcal{S}, g_{\text{LGA}}$]	8	38.0	48.3	21.8	49.3	12.7	46.4	11.6	39.9	0.567	0.783	16.8	11.9	27.9	1.4	32.4
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	8	37.2	47.8	21.4	48.4	12.9	40.9	14.7	33.7	0.806	0.487	17.1	13.2	26.3	2.3	33.5
	EST[$\mathcal{R}$] [51]	32	8.5	13.0	2.6	25.2	2.7	97.9	0.4	97.3	0.848	1.576	1.4	0.0	4.0	0.0	5.3
	ACT[$\mathcal{R}$] [53]	32	27.5	38.2	12.2	43.0	15.5	37.7	15.1	32.2	0.472	0.821	11.1	9.4	14.9	1.0	33.9
	HM[$\mathcal{S}, g_{\text{LGA}}$]	32	36.5	46.7	21.0	48.6	14.7	39.6	15.6	34.2	0.523	0.736	16.5	15.0	26.7	2.9	35.9
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	32	34.9	45.0	19.8	47.1	15.5	37.7	16.6	30.9	0.753	0.506	17.9	16.7	27.3	2.9	36.0
CARS	N/A[$\mathcal{R}$]	N/A	62.5	74.0	23.8	57.0	0.2	100.0	0.1	99.6	0.874	1.816	0.0	0.0	13.4	0.0	3.6
	EST[$\mathcal{R}$] [51]	8	57.1	68.4	30.3	47.7	0.1	99.9	0.1	98.1	0.902	1.681	0.7	0.2	15.4	0.0	4.4
	ACT[$\mathcal{R}$] [53]	8	46.8	58.0	23.4	45.5	19.3	33.1	20.3	32.3	0.413	0.760	18.4	15.0	28.6	1.2	39.8
	HM[$\mathcal{S}, g_{\text{LGA}}$]	8	63.2	73.7	36.8	53.5	15.3	32.0	17.9	33.9	0.463	0.653	23.4	28.5	44.6	5.8	42.4
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	8	61.7	72.6	35.5	51.8	21.0	23.3	23.1	22.2	0.698	0.415	31.2	38.0	47.8	9.6	47.9
	EST[$\mathcal{R}$] [51]	32	30.7	41.0	5.6	31.8	1.2	98.1	0.4	91.8	0.880	1.281	2.9	0.7	8.2	0.0	7.3
	ACT[$\mathcal{R}$] [53]	32	43.4	54.6	11.8	42.9	18.0	32.3	17.5	30.5	0.383	0.763	16.3	15.3	20.7	1.6	38.6
	HM[$\mathcal{S}, g_{\text{LGA}}$]	32	62.3	72.5	35.3	52.7	17.4	28.2	18.2	28.8	0.426	0.613	27.1	30.7	42.3	7.9	44.9
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	32	60.2	71.6	33.9	51.2	19.3	25.9	19.6	25.7	0.650	0.446	30.3	36.7	46.0	8.8	46.0
SOP	N/A[$\mathcal{R}$]	N/A	62.9	68.5	39.2	87.4	0.1	99.3	0.2	99.1	0.845	1.685	0.0	0.0	6.3	0.0	4.0
	EST[$\mathcal{R}$] [51]	8	52.7	58.5	30.1	85.7	6.4	69.7	3.9	64.6	0.611	1.053	3.8	2.2	10.2	1.3	19.0
	ACT[$\mathcal{R}$] [53]	8	45.3	50.6	24.1	84.7	24.8	10.7	25.4	8.2	0.321	0.485	15.4	17.7	25.1	11.3	49.5
	HM[$\mathcal{S}, g_{\text{LGA}}$]	8	49.0	54.1	26.4	85.0	29.9	4.7	31.6	3.6	0.455	0.283	39.3	40.9	38.8	43.0	61.7
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	8	48.3	53.4	25.7	84.9	32.5	4.8	32.4	3.5	0.586	0.239	38.6	39.8	38.3	44.5	61.2
	EST[$\mathcal{R}$] [51]	32	46.0	51.4	24.5	84.7	12.5	43.6	10.6	34.8	0.468	0.830	9.6	7.2	17.3	3.8	31.7
	ACT[$\mathcal{R}$] [53]	32	47.5	52.6	25.5	84.9	24.1	10.5	22.7	9.4	0.253	0.532	21.2	21.6	27.8	15.3	50.8
	HM[$\mathcal{S}, g_{\text{LGA}}$]	32	47.7	52.7	25.3	84.8	30.6	4.7	31.2	3.5	0.466	0.266	38.6	40.3	38.6	44.3	61.8
	HM[$\mathcal{S}, g_{\text{LGA}}$]&ICS	32	46.8	51.7	24.5	84.7	32.0	4.2	33.7	3.0	0.606	0.207	39.1	39.8	37.9	45.6	61.6

Table 6. Comparison of our defense with the state-of-the-art methods on commonly used DML datasets.

Conclusions

- Defense for DML is important.
- HM: flexible and efficient tool. Contribution 1
- LGA: balancing training objectives. Contribution 2
- ICS: further improve robustness. Contribution 3
- Outperforms SoTA by a margin.

Thanks!