IEEE International Conference on Image Processing
27 - 30 October 2024 | Abu Dhabi, UAE

ADVERSARIAL ROBUSTNESS FOR DEEP METRIC LEARNING

Ezgi Paket, İnci M. Baytaş

Boğaziçi University
İstanbul / Türkiye

2024

Deep Metric Learning

- Learning a non-linear projection to a new space
- Minimizing distance between semantically similar samples
- Maximizing the distance between dissimilar samples

LIMITATION

- Vulnerability to human-imperceptible perturbations (Adversarial Attacks) [1]

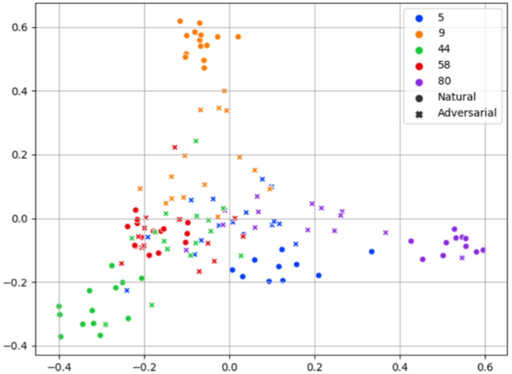


Figure: T-SNE of natural and adversarial embeddings for a sample data from CUB200-2011 dataset. It illustrates that adversarial samples move away from their natural counterparts, while reducing the distance between the adversarial and natural samples from different categories

[1] Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks", ArXiv:1312.6199 [cs], 2013.

Adversarial Attacks

- OBJECT CLASSIFICATION
- IMAGE CLUSTERING
- IMAGE RETRIEVAL
- FACE RECOGNITION

Attack Types

- White-Box
- Gray-Box
- Black-Box

Adversarial Defenses in DML Literature

Adversarial training from scratch:

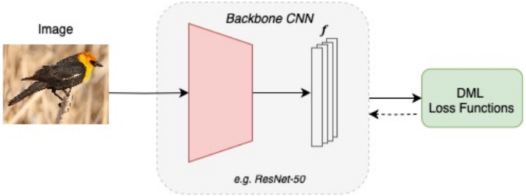


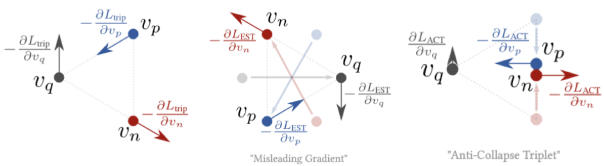
Figure: Backbone architecture with DML loss functions.

Triplet Loss Adversarial (TLA) [2]

Figure: Illustration of triplet loss for TLA.

- PGD attack to cross-entropy loss
- Regularizing cross-entropy with triplet loss

Anti-Collapse Triplet (ACT) [3]



Before attack | After attack | After attack

Figure: Misleading gradients in arbitrary attacks vs. gradient direction of Anti-Collapse Triplet (ACT).

[2] Mao, C., Z. Zhong, J. Yang, C. Vondrick and B. Ray, "Metric learning for adversarial robustness", Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, pp. 480–491, 2019.

[3] Zhou, M., L. Wang, Z. Niu, Q. Zhang, N. Zheng and G. Hua, "Adversarial attack and defense in deep ranking", ArXiv:2106.03614 [cs], 2021.

Adversarial Defenses in DML Literature

<u>Adversarial training via fine-tuning of pretrained networks:</u>

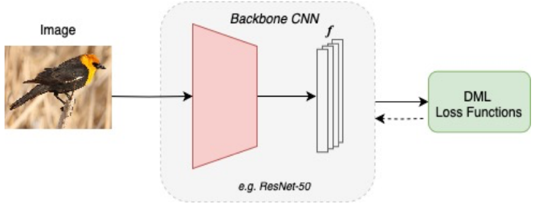


Figure: Backbone architecture with DML loss functions.

Robust deep metric learning via fine-tuning [4]

- PGD attack to contrastive and triplet loss
- Training with contrastive and triplet loss

Adversarial Deep Metric Learning (ADML) [5]

- PGD attack to alignment loss
- Training with alignment & uniformity loss

[4] Panum, T. K., Z. Wang, P. Kan, E. Fernandes and S. Jha, "Exploring adversarial robustness of deep metric learning", ArXiv:2102.07265 [cs], 2021.
[5] Wu, Y. and H. Huang, "Understanding Metric Learning on Unit Hyper-sphere and Generating Better Examples for Adversarial Training", 2022, https://openreview.net/forum?id=DkeCkhLIVGZ .

Proposed Method

Contributions in this study:

- A lightweight, robust metric learning (RML) approach without generating adversarial samples during training
- Reduced training complexity and time
- Maintained SOTA performance on the natural samples
- Does not depend on specific architectures

Robust Metric Learning

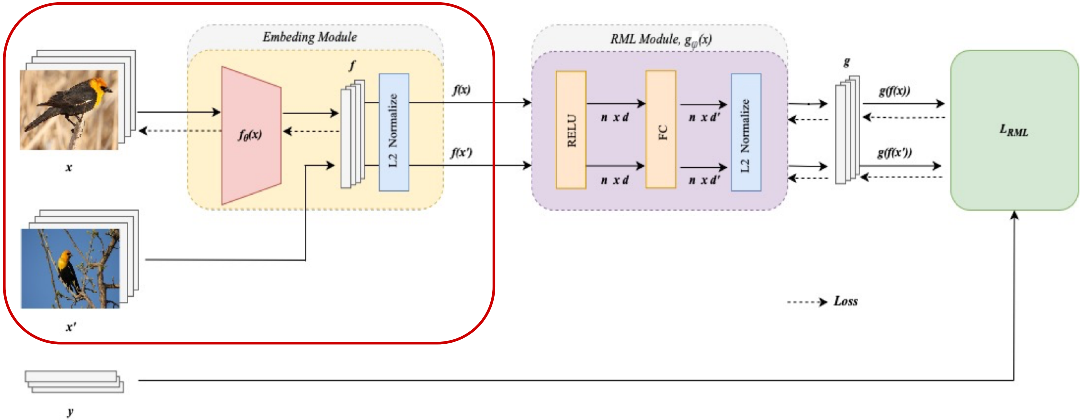


Figure: Proposed robust deep metric learning model. Embeddings of natural, $f(x;\ \theta) \epsilon \mathbb{R}^d$, and adversarial images, $f(x';\ \theta) \epsilon \mathbb{R}^d$, are extracted using embedding module. The embedding module is frozen, while the metric learning module is training. The outputs of the metric learning module, $g(f(x;\ \theta); \phi) \epsilon \mathbb{R}^{d'}$, and $g(f(x';\ \theta); \phi) \epsilon \mathbb{R}^{d'}$, are provided to related loss function.

Method - Embedding Module

<u>Outputs of Embedding Module:</u>

- Embeddings of natural images
- Embeddings of adversarial images

Step 1: Fine-tuning of pre-trained architectures using cross-entropy loss with **only natural** samples.

Step 2: Adversarial attack generation:

Require: Natural data: $\mathcal{D}_{\text{nat}} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$,
adversarial data: $\mathcal{D}_{\text{adv}} = \{(\mathbf{x}^1_{adv}, y_1), \cdots, (\mathbf{x}^N_{adv}, y_N)\}$
Ensure:
for mini-batch $\{\mathbf{x}_i, \boldsymbol{y}_i\}_{i=1}^n \sim$ D **do**
  $\mathbf{x}_{\text{a}}, \mathbf{x}_{\text{p}}, \mathbf{x}_{\text{n}} \leftarrow$ anchor, positive, negative images
  $\mathbf{x}_{\text{adv}} \leftarrow \mathbf{x}_{\text{a}}$
  for m = 1, ... , M **do**
    $\mathbf{h}_{\text{adv}}, \mathbf{h}_{\text{p}}, \mathbf{h}_{\text{n}} \leftarrow f(\mathbf{x}_{\text{adv}}; \theta), f(\mathbf{x}_{\text{p}}; \theta), f(\mathbf{x}_{\text{n}}; \theta)$
    $\boxed{\mathbf{x}_{\text{adv}} = (\mathbf{x}_{\text{adv}} + \gamma \cdot \text{sign}(\nabla_{x_{adv}} \mathcal{L}_{\text{Triplet}}(\mathbf{h}_{\text{adv}}, \mathbf{h}_{\text{p}}, \mathbf{h}_{\text{n}})))}$
    $\boldsymbol{\delta} = \max(\min(\mathbf{x}_{\text{adv}} - \mathbf{x}_{\text{a}}, \epsilon), -\epsilon)$
    $\mathbf{x}_{\text{adv}} = \mathbf{x}_{\text{a}} + \boldsymbol{\delta}$
  end for
end for

Method

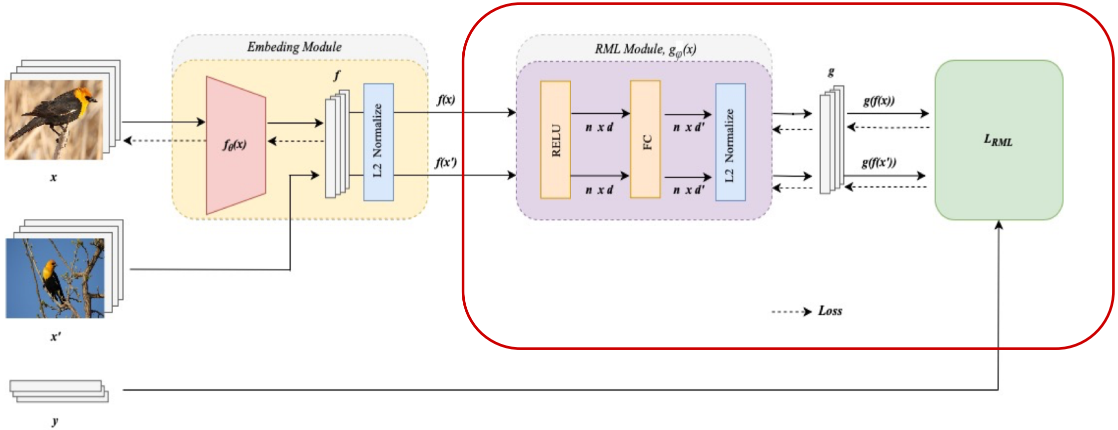


Figure: Proposed robust deep metric learning model. Embeddings of natural, $f(x;\ \theta) \epsilon \mathbb{R}^d$, and adversarial images, $f(x';\ \theta) \epsilon \mathbb{R}^d$, are extracted using embedding module. The embedding module is frozen, while the metric learning module is training. The outputs of the metric learning module, $g(f(x;\ \theta); \phi) \epsilon \mathbb{R}^{d'}$, and $g(f(x';\ \theta); \phi) \epsilon \mathbb{R}^{d'}$, are provided to related loss function.

Robust Metric Learning (RML)

Require: Natural data: $\mathcal{D}_{\text{nat}} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$,

adversarial data: $\mathcal{D}_{\text{adv}}$,

data embeddings: h

Ensure:

Fine-tuning: $f \leftarrow \text{FinetuneResNet}(D_{nat})$

Adversarial attack generation: $\mathcal{D}_{\text{adv}} \leftarrow \text{PGDAttack}(f(D_{\text{nat}}, \theta))$

Embedding module: $\mathbf{h}_{\text{nat}}, \mathbf{h}_{\text{adv}} \leftarrow \text{GetEmbedding}(D_{\text{nat}}, D_{\text{adv}})$

Triplet sampling: anchor: $\mathbf{h}_{\text{adv}}$, positive: $\mathbf{h}_{\text{p}}$, negative: $\mathbf{h}_{\text{n}}$

for t = 1, ... , T **do**

  for mini-batch $\{\mathbf{x}_i, \boldsymbol{y}_i\}_{i=1}^{n} \sim \text{D}$ **do**

    model update:

    $\phi \leftarrow \phi - \tau \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \mathcal{L}_{\text{RML}}\left(g\left(\mathbf{h}_{\text{adv}}^{\text{i}}; \phi\right), g\left(\mathbf{h}_{\text{p}}^{\text{i}}; \phi\right), g\left(\mathbf{h}_{\text{n}}^{\text{i}}; \phi\right)\right)$

  end for

end for

Figure: Adversarial Metric Learning Framework.

Robust Metric Learning (RML)

$$\mathcal{L}^{RML}_{Contrastive}\ (h_{adv}, h_{comp}; \phi)$$

$$= \frac{1}{N}\ \sum_{i=1}^{N}[\ (1-Y)\,\frac{1}{2}\mathrm{D}(g(h^{i}_{adv};\phi),\ g(h^{i}_{comp};\phi))^{2} + \mathrm{Y}\ \frac{1}{2}\ (m\ -\mathrm{D}(g(h^{i}_{adv};\phi),\ g\big(h^{i}_{comp};\phi\big))^{2}\ )]$$

$$\mathcal{L}^{RML}_{Triplet}\ (h_{adv}, h_{p}, h_{n}; \phi)$$

$$= \frac{1}{N}\ \sum_{i=1}^{N}[\ \mathrm{D}\ (g(h^{i}_{adv};\phi),\ g(h^{i}_{p};\phi))\ -\mathrm{D}(g(h^{i}_{adv};\phi),\ g\big(h^{i}_{n};\phi\big)) + m]$$

$$\mathcal{L}^{RML}_{Angular}\ (h_{adv}, h_{p}, h_{n}; \phi)$$

$$f_{adv,p,n} = 4\ tan^{2}\alpha(g\left(h_{adv};\phi\right),\ g\big(h_{p};\phi\big)\Big)^{T} g(h_{n};\phi) - 2(1+tan^{2}\alpha)\ g\left(h_{adv};\phi\right)^{\mathsf{T}}\ g\big(h_{p};\phi\big)\Big)$$

$\mathcal{L}$: Loss function
h_{comp}: Comparison embedding
h_{adv} : Adversarial anchor embedding
h_{p}: Natural positive embedding
h_{n}: Natural negative embedding
D:Distance

ϕ: Network parameters
$g(.)$: RML model
m :Pre-determined margin
Y: Positive/negative label
α: Target angle

Experiments: Datasets

Figure: CUB200-2011.

CUB200-2011 [6]

consists of 200 bird classes with **11,788** images in total. While training data contains the first 100 classes with 5,864 images, the test set has the other 100 classes with 5,924 images.

CARS196 [7]

includes **16,185** car images from 196 different classes. While the train set has the first 98 types of cars with 8,144 images, the test set includes the last 98 classes with 8,041 images.

Figure: CARS196.

Figure: SOP.

Stanford Online Products (SOP) [8]

SOP dataset has 22,634 classes with **120,053** images. It includes 59,551 images from 11,318 classes for training and 60,502 images from the remaining 11,157 classes for testing.

[6] Wah, C., S. Branson, P. Welinder, P. Perona and S. Belongie, Caltech-ucsd birds 200, Tech. Rep. CNS-TR-2011-001, California Institute of Technology, California, CA, USA, 2011.
[7] Krause, J., M. Stark, J. Deng and L. Fei-Fei, "3d object representations for fine-grained categorization", Proceedings of the IEEE international conference on computer vision workshops, Sydney, Australia, pp. 554–561, 2013.
[8] Song, H. O., Y. Xiang, S. Jegelka and S. Savarese, "Deep metric learning via lifted structured feature embedding", Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, pp. 4004–4012, 2016.

Quantitative Results

ResNet50

		CUB200-2011					CARS196					SOP				
Model	Dim	NMI	R@1	R@2	R@4	R@8	NMI	R@1	R@2	R@4	R@8	NMI	R@1	R@10	R@100	R@1000
Natural Samples																
ResNet-50 [11]	2048	57.8	47.5	61.6	73.0	83.9	42.0	44.2	56.7	68.1	78.6	86.2	54.3	70.6	83.7	94.5
FT ResNet-50 [29]	2048	**71.4**	**87.7**	**90.2**	**93.2**	**95.3**	**74.2**	**96.1**	**96.2**	**97.6**	**98.6**	**94.2**	**91.4**	**94.8**	**97.3**	**99.0**
EARDML$_{\text{Contrastive}}$ [1]	128	-	58.2	-	-	-	-	72.1	-	-	-	-	66.7	-	-	-
EARDML$_{\text{Triplet}}$ [1]	128	-	53.4	-	-	-	-	71.9	-	-	-	-	64.0	-	-	-
RML$_{\text{Contrastive}}$	512	64.5	85.1	87.7	90.6	93.6	68.8	93.3	95.0	96.7	98.0	93.2	88.6	92.4	95.6	98.0
RML$_{\text{Angular}}$	512	70.1	87.0	89.6	92.5	94.9	56.5	85.4	87.4	90.4	93.6	91.3	82.8	87.5	92.4	96.7
RML$_{\text{Triplet}}$	512	68.9	87.0	89.4	92.2	94.7	72.5	94.0	95.4	97.1	98.1	93.4	89.6	93.2	96.3	98.4
PGD-5 ($\epsilon = 0.01$)																
ResNet-50 [11]	2048	27.5	12.5	20.1	29.6	42.3	18.5	10.8	16.2	23.8	33.4	80.8	21.2	35.3	54.8	78.3
FT ResNet-50 [29]	2048	23.8	17.1	23.1	31.4	42.3	19.4	27.5	35.2	44.8	56.1	86.4	69.6	77.4	84.4	90.9
EARDML$_{\text{Contrastive}}$ [1]	128	-	20.3	-	-	-	-	35.7	-	-	-	-	53.6	-	-	-
EARDML$_{\text{Triplet}}$ [1]	128	-	16.9	-	-	-	-	36.2	-	-	-	-	39.3	-	-	-
RML$_{\text{Contrastive}}$	512	**27.9**	19.7	26.4	34.2	45.3	27.5	38.1	47.4	57.2	67.7	**89.1**	**77.7**	**83.1**	**88.3**	93.4
RML$_{\text{Angular}}$	512	24.0	17.5	23.4	31.0	42.1	**27.7**	15.8	22.8	32.3	44.3	87.5	67.7	77.7	85.6	93.0
RML$_{\text{Triplet}}$	512	27.8	**22.6**	**29.7**	**38.9**	**49.9**	27.4	**39.5**	**48.1**	**58.1**	**69.0**	88.1	75.2	81.9	87.7	**93.5**

Table: Natural and adversarial performances of robust metric learning module trained ResNet-50 embeddings.

Quantitative Results

ResNet18

		CUB200-2011					CARS196					SOP				
Model	Dim	NMI	R@1	R@2	R@4	R@8	NMI	R@1	R@2	R@4	R@8	NMI	R@1	R@10	R@100	R@1000
Natural Samples																
FT ResNet-18 [29]	512	**66.3**	**85.1**	**87.8**	**91.2**	**94.3**	**67.2**	**92.4**	**94.0**	**96.2**	**97.8**	**93.7**	**89.9**	**93.7**	**97.0**	**98.8**
$\text{RML}_{\text{Contrastive}}$	1024	61.3	82.3	84.7	88.6	92.4	61.7	89.8	92.0	94.4	96.5	93.1	88.4	92.4	95.9	98.2
$\text{RML}_{\text{Angular}}$	1024	58.9	81.0	83.7	87.7	91.4	56.5	85.7	88.0	91.2	93.8	92.3	86.6	90.7	94.4	97.2
$\text{RML}_{\text{Triplet}}$	1024	61.0	81.9	84.5	88.6	92.6	61.9	89.5	91.5	94.1	96.1	93.0	88.2	92.3	95.8	98.2
PGD-5 ($\epsilon = 0.01$)																
FT ResNet-18 [29]	512	22.1	15.5	21.4	29.4	39.5	17.7	12.9	19.5	28.6	40.9	84.4	53.3	64.9	76.1	86.7
$\text{RML}_{\text{Contrastive}}$	1024	**26.1**	**21.6**	**28.9**	**38.0**	**48.4**	**22.1**	**16.5**	23.8	33.3	45.5	85.3	58.1	69.8	79.8	88.6
$\text{RML}_{\text{Angular}}$	1024	22.6	12.9	17.9	24.8	34.3	17.9	8.4	12.9	19.9	29.2	82.7	46.2	57.8	68.7	80.8
$\text{RML}_{\text{Triplet}}$	1024	25.6	20.3	27.0	35.5	46.6	**22.1**	16.1	**23.9**	**33.5**	**45.6**	**85.7**	**58.7**	**70.9**	**81.1**	**90.2**

Table: Natural and adversarial performances of robust metric learning module trained with ResNet-18 embeddings.

Quantitative Results

	CUB200-2011		CARS196		SOP	
Models	R@1	NMI	R@1	NMI	R@1	NMI
ADML + T [33]	11.58	25.3	25.4	21.2	10.7	80.2
ADML + A [33]	17.4	**29.2**	40.0	26.1	14.0	80.4
ADML + U [33]	15.1	27.9	33.1	24.5	11.3	80.3
RML	**24.0**	27.4	**50.9**	**35.3**	**73.0**	**88.0**

Table: Adversarial robustness of different approaches including the proposed RML against adversarial samples synthesized by attacking alignment loss.

Quantitative Results

Operations	Required Time (minute)
FT ResNet-50	58.0
Attack Generation	12.7
Adversarial Metric Learning	2.7

Table: Training time analysis for the proposed approach. Training time of the first epoch is measured as 0.58 minutes, and it is multiplied by 100 epochs for the fine-tuning of the pre-trained ResNet-50 model naturally. Adversarial attack generation is completed in 12.7 minutes. Robust metric learning is applied for 2.7 minutes.

Methods	Required Time (hour)
Adversarial Training	21.7
Our Approach	1.2

Table: Training time comparisons. Training time for an epoch is calculated as 13 minutes and it is multiplied by 100 epochs for an adversarial training.

Qualitative Results

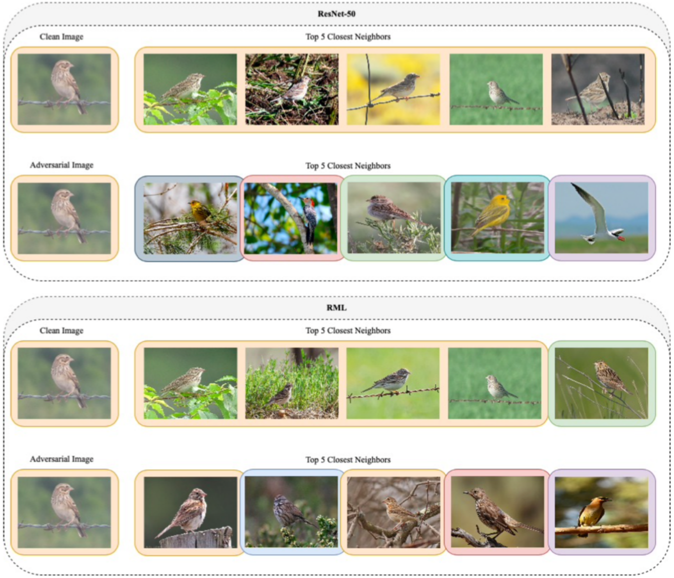


Figure: Top 5 nearest neighbors comparisons of original ResNet-50 and RML embeddings in natural and adversarial settings for CUB dataset.

Conclusion & Future Work

Conclusion

- Adversarial samples are generated once and saved to be utilized in the following metric learning module in a black-box manner. Thus, the training time and complexity are reduced while improving and sometimes preserving the state-of-the-art robustness of models.
- The proposed lightweight metric learning module maintains natural performances similar to original embeddings.
- The robust metric learning module is adaptable to different deep backbone architectures.

Future Work

- Model performances can be tested under various attack configurations.
- Exploring proper data augmentation techniques for each dataset can be the further research area to extend this study.

THANK YOU

For further questions:

Ezgi Paket

ezgi.paket.2024@alumni.boun.edu.tr

İnci M. Baytaş

inci.baytas@bogazici.edu.tr