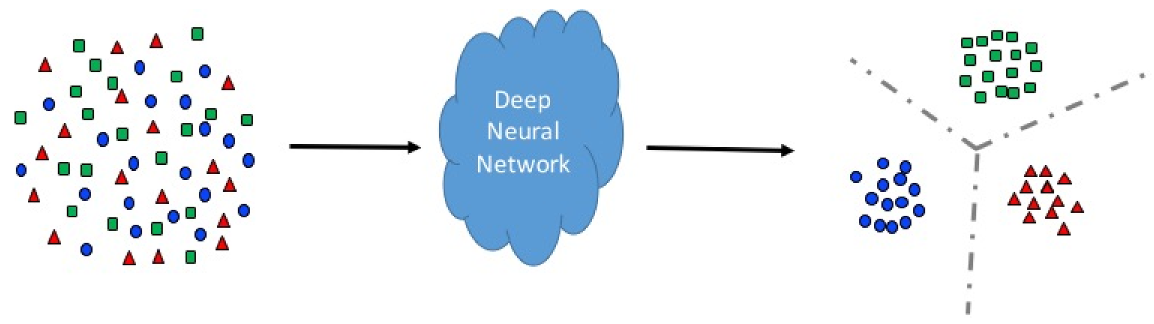
A Simple and Effective Framework for Pairwise Deep Metric Learning

Qi Qi, Yan Yan, Zixuan Wu Xiaoyu Wang, Tianbao Yang
qi-qi@uiowa.edu, yanyan.tju@gmail.com
University of Iowa.

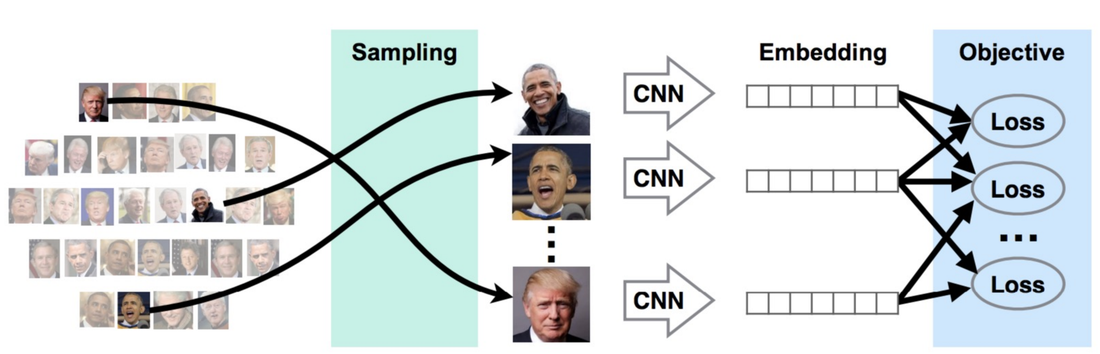
Introduction

Deep Metric Learning:

- Task: Learning a metric to measure the distance between pairs by training a deep neural network.
- Goal: Euclidean distance of pairs from the same class shall be small, while pairs from different classes shall be large.



Overview of Training Process:



- Pair-based Losses: given two examples $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j)$, deep neural network is parametrized as θ:

$$\ell_{ij}(\Theta) = \ell(f(\mathbf{x}_i;\theta), f(\mathbf{x}_j;\theta); y_{ij}) \quad (1)$$

where $y_{ij}=1$ if $y_i = y_j$, and $y_{ij}=0$ if $y_i \neq y_j$. $f(\cdot,\theta)$ is the output of the neural network.

Optimization

A mini-batch of examples denoted by $\{\mathbf{x}_1, ..., \mathbf{x}_B\}$, B is the batch size. B^2 pairs are constructed between this samples. The naive approach (most common) for DML is minimizing average loss function in terms of θ within a batch:

$$\mathcal{L}_{\text{avg}}(\theta) = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} l_{ij}(\theta)$$

- **Design More Complicated Losses**:
  - Lifted-Structure(LS) [1] loss:

$$\mathcal{L}_{LS} = \sum_{i=1}^{B}[\log\sum_{k\in P_i} e^{\lambda - S_{ik}} + \log\sum_{k\in N_i} e^{S_{ik}-\lambda}]_+, \quad (2)$$

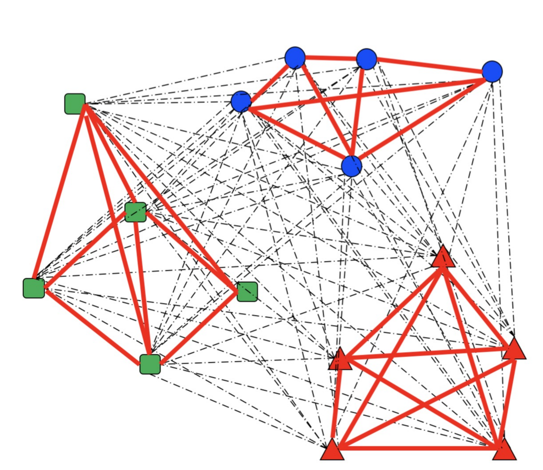
  - Multi-Similarity(MS) [2] loss:

$$\mathcal{L}_{MS} = \frac{1}{B}\sum_{i=1}^{B}\{\frac{1}{\alpha}\log[1+\sum_{k\in P_i} e^{-\alpha(S_{ik}-\lambda)}] + \frac{1}{\beta}\log[1+\sum_{k\in N_i} e^{\beta(S_{ik}-\lambda)}]\} \quad (3)$$

  where λ, α, β are hyper-parameters. $S_{ij} =< f(\mathbf{x}_i;\theta), f(\mathbf{x}_j;\theta) >$ denotes the similarity of the two samples in the embedding space.
- **Mining Strategy (Sampling)**:
  - Hard (Seimi-Hard): Select hard negative pairs whose distance is smaller than that between the positive pairs.
  - Distance Weighted Sampling (DWS): Negative pairs sampled according to their distance distribution within a batch.

Deficiency:

- Losses are more and more complicated but hard to understand, and also fail to explain why its effectiveness.
- Heuristic and lack of theoretical guarantee.
- Fail to address the most fundamental challenge: Pair Imbalance



- represent positive pairs, - - represent negative pairs.

Contributions

- We proposed a general DRO framework for DML. Theoretical justification of the proposed framework is provided from the perspective of advance learning theories.
- The proposed general DRO framework can recover SOTA complicated pair-based losses: MS Loss and LS Loss by specifying different uncertainty sets.
- More effective solutions has been provided under DRO framework for tackling DML. Experimental results show that our proposed variants of DRO framework outperform SOTA methods on several benchmark datasets

General DRO-based Framework

- $$\mathcal{L}(\theta) = \max_{\mathbf{p}\in\mathcal{U}} g(\theta, \mathbf{p}) := \sum_{i=1}^{B}\sum_{j=1}^{B} p_{ij} l_{ij}(\theta), \quad (4)$$

  where $\mathbf{p} \in \mathbb{R}_+^{B^2}$ is a non-negative vector with each element p_{ij} representing a weight (sampling probability) for an individual pair. $\mathcal{U} \subseteq \mathbb{R}^{B^2}$ denotes the decision set of $\mathbf{p}$.
- $\mathcal{L}(\theta)$ is more robust to pair imbalance than $\mathcal{L}_{avg}$.
- Theoretical analysis in [3, 4] verified that $\mathcal{L}(\theta)$ is a better approximation than $\mathcal{L}_{avg}(\theta)$ for $\mathbf{E}[\ell(\theta)]$.
- LS loss and MS loss can be recovered by setting $\mathcal{U}$.

Theoretical Guarantees

Let $Z = \{Z_1, ..., Z_n\}$ be i.i.d. random losses taking values in $[M_0, M_1]$ where $M = M_1 - M_0$. Suppose $\hat{\mathbf{p}}_n = (1/n, \ldots, 1/n)$ is the empirical distribution, $\mathcal{U}_\phi = \{\sum_i p_i = 1, p_i \geq 0, D_\phi(\mathbf{p}\|\hat{\mathbf{p}}_n) \leq \frac{\rho}{n}\}$. Denote the empirical variance of $Z_1, \ldots, Z_n$ by $\text{Var}_n(Z)$ and fix $\rho \geq 0$. If $n \geq \max\{\frac{24\rho}{\text{Var}(Z)}, \frac{16}{\text{Var}(Z)}, 1\}M^2$, then

$$\sup_{\mathbf{p}\in\mathcal{U}_\phi}\sum_{i=1}^{n} p_i Z_i = \frac{1}{n}\sum_{i=1}^{n} Z_i + \sqrt{\frac{2\rho\text{Var}_n(Z)}{n}}.$$

Three Variants DRO for DML

For each $\mathbf{x}_i$ serve as an anchor in a given mini-batch whose size is B, $\mathbf{P}_i = \{j|y_{ij} = 1, j \in [B]\}$ and $\mathbf{N}_i = \{j|y_{ij} = 0, j \in [B]\}$ denote the index sets of positive and negative pairs, respectively. $\mathbf{P} = \cup_{i=1}^{B}\mathbf{P}_i$ and $N = \cup_{i=1}^{B}\mathbf{N}_i$. Three variants of general framework with different uncertainty set $\mathcal{U}$ is defined as follows:

- DRO-TopK:

$$\max_{\mathbf{p}} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i\cup\mathbf{N}_i} p_{ij} l_{ij}(\theta)$$

$$\text{s.t.} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i\cup\mathbf{N}_i} p_{ij} = 1, 0 \leq p_{ij} \leq 1/K,$$

- DRO-TopK-PN:

$$\max_{\mathbf{p}\in\{0,1\}^{P+N}} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i\cup\mathbf{N}_i} p_{ij} l_{ij}(\theta)$$

$$\text{s.t.} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i} p_{ij} \leq \frac{K}{2}, \sum_{i=1}^{B}\sum_{j\in\mathbf{N}_i} p_{ij} \leq \frac{K}{2}.$$

- DRO-KL:

$$\max_{\mathbf{p}\in\mathbb{R}_+^{P+N}} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i\cup\mathbf{N}_i} p_{ij} l_{ij}(\theta) - \gamma D_{KL}(\mathbf{p}||\frac{\mathbf{1}}{P+N}),$$

$$\text{s.t.} \sum_{i=1}^{B}\sum_{j\in\mathbf{P}_i\cup\mathbf{N}_i} p_{ij} = 1,$$

  where $\gamma > 0$ is a hyper-parameter and D_{KL} denotes the KL divergence between two distributions.
- Close-form $\mathbf{p}$ can be derived using KKT-Condition.

Empirical Studies

Datasets

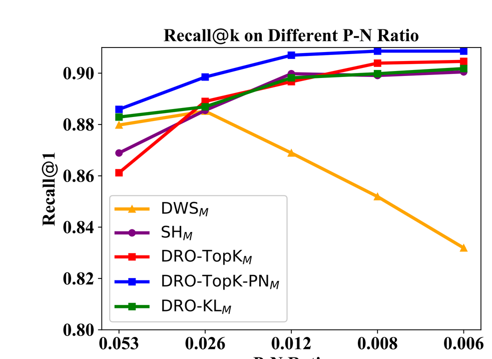
- Three Benchmark Datasets

Data Sets	# of training	# of testing	# of Classes
Cub-200-2011	5864	5924	200
Cars-196	8054	8131	196
In-shop	14,218	12,612	7,970

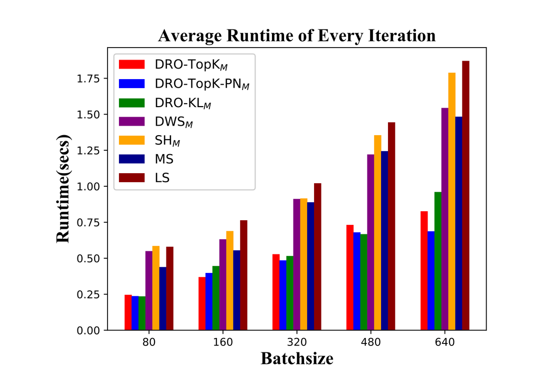
- Evaluation Metric: Recall@k
- Margin Loss:

$$\ell_{ij}(\theta) = [\alpha + y_{ij}(\lambda - S_{ij})]_+$$

Imbalance and Runtime



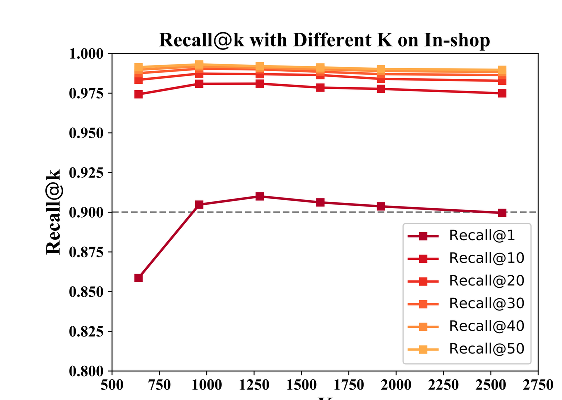
(a) Recall vs Imbalance Ratio



(b) Average running time of every iteration

Figure: Quantitive comparison with SOTA pair mining and complicated losses

Sensitivity of K



Recover of LS and MS

Recall@K(%)	1	10	20	30	40	50
MS	79.8	94.9	96.8	97.6	97.9	98.3
LS	82.6	94.1	95.6	96.4	96.9	97.4
DRO-KL-G-$\gamma = 1$	**84.8**	**95.9**	**97.3**	**97.9**	**98.2**	**98.5**
DRO-KL-G-$\gamma = 0.1$	**85.1**	**96.1**	**97.5**	**98.0**	**98.3**	**98.5**
DRO-KL-G-$\gamma = 0.01$	**85.8**	**96.2**	**97.9**	**97.8**	**98.2**	**98.4**
DRO-KL-G-$\gamma = 0.001$	**85.7**	**96.1**	**97.4**	**97.9**	**98.2**	**98.5**

Table: Recover of MS loss and LS loss on In-Shop

SOTA Quantitive Results

Recall@K	1	10	20	30	40	50
FashionNet	53.7	73.0	76.0	77.0	79.0	80.0
HDC	62.1	84.9	89.0	91.2	92.3	93.1
HDL	80.9	94.3	95.8	97.2	97.4	97.8
ABIER	83.1	95.1	96.9	97.5	97.8	98.0
ABE	87.3	96.7	97.9	98.2	98.5	98.7
MS	89.7	97.9	98.5	**98.8**	**99.1**	**99.2**
DRO-TopK$_M$(Ours)	**91.0**	**98.1**	**98.7**	**99.0**	**99.1**	**99.2**
DRO-TopK$_B$(Ours)	**90.7**	97.7	98.4	**98.8**	99.0	99.1
DRO-TopK-PN$_M$(Ours)	**91.3**	**98.0**	**98.7**	**98.9**	**99.1**	**99.2**
DRO-TopK-PN$_B$(Ours)	**91.1**	**98.1**	**98.6**	**98.8**	99.0	**99.2**
DRO-KL$_M$(Ours)	**90.8**	**98.0**	**98.6**	**99.0**	**99.1**	**99.2**

Table: Recall@k on In-Shop

References

[1] Oh Song, Hyun, et al. Deep metric learning via lifted structured feature embedding. *CVPR* 2016.

[2] Wang, Xun, et al. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. *CVPR*, 2019.

[3] Namkoong, Hongseok, and John C. Duchi. Variance-based regularization with convex objectives. *NIPS*, 2017

[4] Maurer, Andreas, and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.