Ranked List Loss for Deep Metric Learning

Xinshao Wang, Yang Hua, Elyor Kodirov,

Guosheng Hu, Romain Garnier, and

Neil M. Robertson

CVPR 2019, Poster

Background

- What is deep metric learning (DML)?
  - Objective: a deep embedding space such that relative locations of input samples are based on their semantic similarities.
  - Key point:

$$\boxed{\text{Input Space}} \xrightarrow{f} \boxed{\text{Embedding Space}}$$

An excerpt of t-SNE plot on CUB-200-2011 test set.

QUEEN'S UNIVERSITY BELFAST anyvision.

Background

- DML is fundamental and learns deep representations => diverse applications
  - Image Retrieval (Lifted Struct, Song et al., CVPR 2016)
  - Person ReID (RCN, McLaughlin et al. CVPR 2016)
  - Clustering (Lifted Struct, Song et al., CVPR 2017)
  - Verification (FaceNet, Schroff et al., CVPR 2015)
  - Few-shot Learning (ProtoNet, Snell et al., NeurIPS 2017)
  - Generative Networks (BourGAN, Xiao et al., NIPS 2018)

Deep Metric Learning versus Few-shot Learning

- Shared goal: learning representations (embedding function)
- Training settings can also be shared
  - Few-shot: learn iteratively on K-shot C-way classification tasks
  - Deep Metric: learn iteratively on mini-batches with C classes x K images per class
- Test settings are different
  - Few-shot: test iteratively on K-shot C-way classification tasks
  - Deep Metric: test on one task with unseen classes generally.

Limitations of Existing Methods

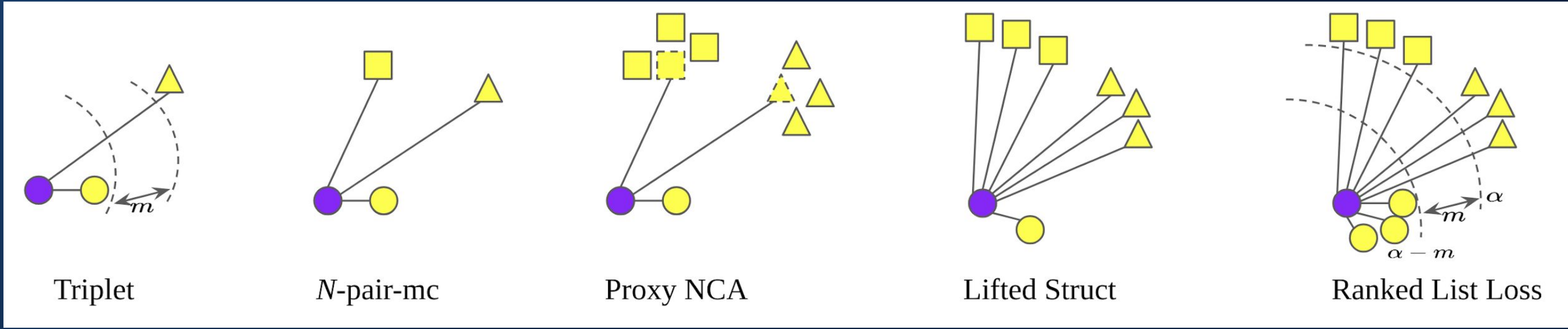
- ❑ Exploiting only a proportion informative examples
  We argue that utilising all non-trivial data points can be better.

- ❑ Tending to shrink class-level distribution into one point, i.e., kill all intraclass variances in the metric space.
  We argue that regularising the metric space by pulling all intraclass examples into a hypersphere is better.

- ❑ Imbalanced number of positive examples and negative ones.

Hypersphere Regulator: Non-excessively Reduce Intraclass Variance

- Illustration of different ranking-motivated structured losses.



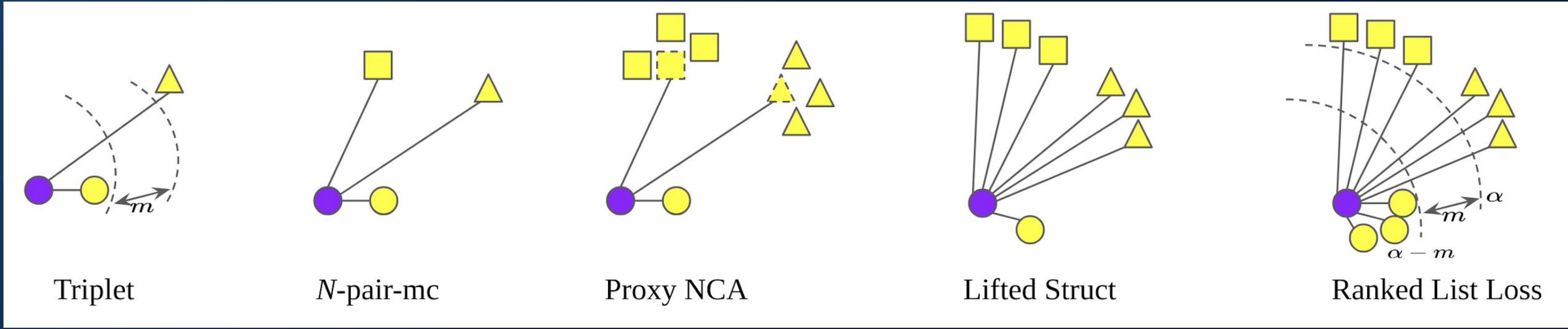
The blue circle is an anchor (query).

Different shapes (circle, triangle and square) represent different classes. For simplicity, only 3 classes are shown.

The intraclass hypersphere diameter: $\alpha - m$

Exploiting All Non-trivial Training Examples

- Illustration of different ranking-motivated structured losses.

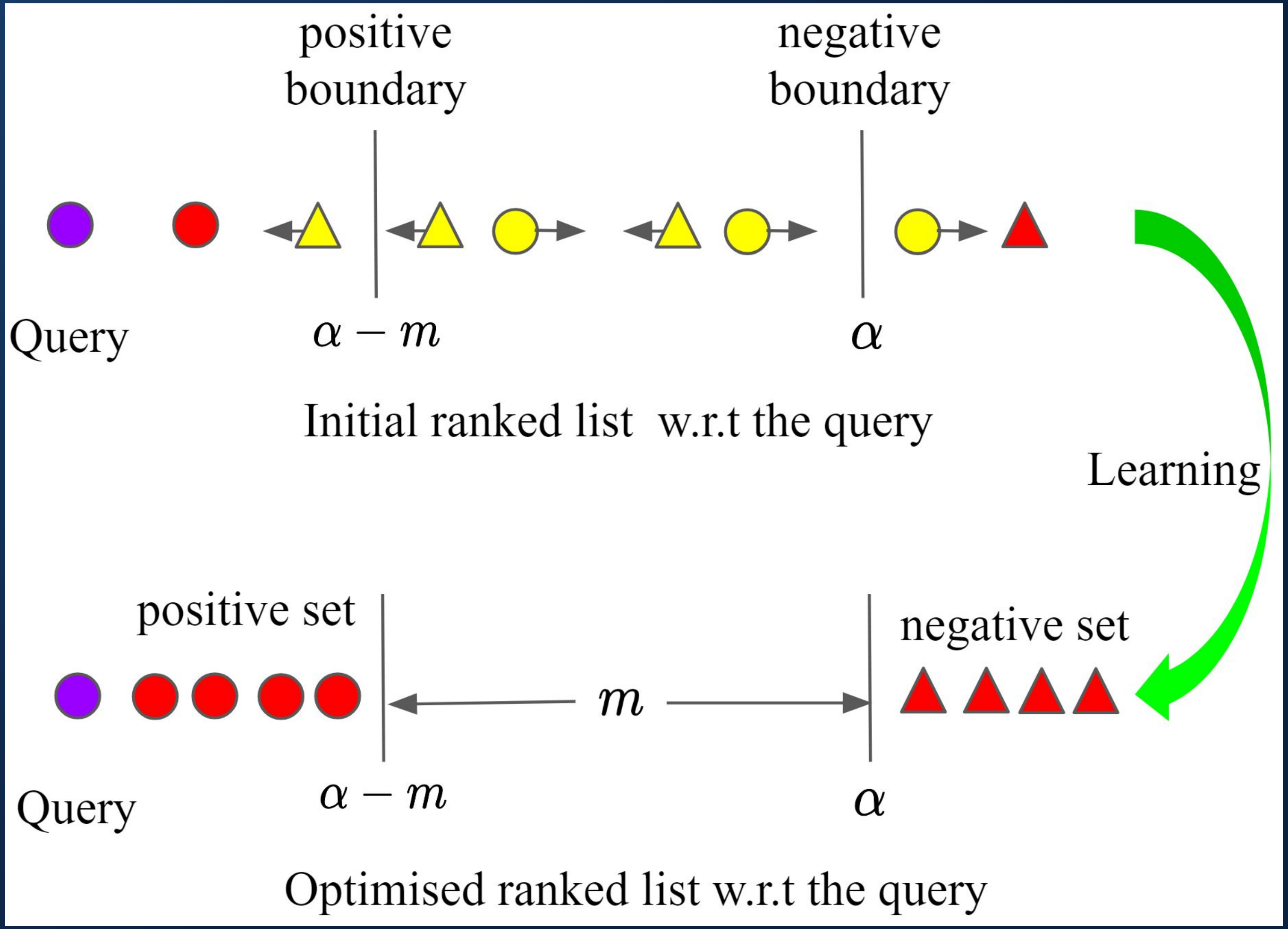


The blue circle is an anchor (query).

Different shapes (circle, triangle and square) represent different classes. For simplicity, only 3 classes are shown.

The intraclass hypersphere diameter: $\alpha - m$

The optimisation objective of RLL



Red shapes: trivial examples
Yellow ones: informative points

Arrow: pull/push effect to query by non-trivial examples in the gallery list

Learning: the query's representation is updated based on a weighted combination of their effects.

Methodology

- Pairwise Constraint

$$L_{\mathrm{m}}(\mathbf{x}_i, \mathbf{x}_j; f) = (1 - y_{ij})[\alpha - d_{ij}]_+ + y_{ij}[d_{ij} - (\alpha - m)]_+,$$

where $y_{ij} = 1$ if $y_i = y_j$, and $y_{ij} = 0$ otherwise. $d_{ij} = ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2$ is the Euclidean distance between two points.

Methodology

- Non-trivial Sample Mining
  - Before Mining

    The positive set with respect to the query x_i^c :

$$\mathrm{P}_{c,i} = \{\mathbf{x}_j^c | j \neq i\}, |\mathbf{P}_{c,i}| = N_c - 1.$$

    The corresponding negative set:

$$\mathrm{N}_{c,i} = \{\mathbf{x}_j^k | k \neq c\}, |\mathbf{N}_{c,i}| = \sum_{k \neq c} N_k.$$

  - After Mining

$$\mathbf{P}_{c,i}^* = \{\mathbf{x}_j^c | j \neq i, d_{ij} > (\alpha - m)\}$$

$$\mathbf{N}_{c,i}^* = \{\mathbf{x}_j^k | k \neq c, d_{ij} < \alpha\}$$

Methodology

- Loss-based Negative Examples Weighting
  - The original weight is one for any representation

$$||\frac{\partial L_{\mathrm{m}}(\mathbf{x}_i, \mathbf{x}_j; f)}{\partial f(\mathbf{x}_j)}||_2 = ||\frac{f(\mathbf{x}_i) - f(\mathbf{x}_j)}{||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2}||_2 = 1.$$

  - Weighting scheme

$$w_{ij} = \exp(T \cdot (\alpha - d_{ij})), \mathbf{x}_j^k \in \mathbf{N}_{c,i}^*.$$

$T = 0$: Treating all non-trivial negative examples equally.

$T = +\infty$: Mining the hardest negative example

Methodology

Imbalanced number of
positive and negative examples

- Optimisation Objective
  - Positive set

$$L_{\mathrm{P}}(\mathbf{x}_i^c; f) = \frac{1}{|\mathbf{P}_{c,i}^*|} \sum_{\mathbf{x}_j^c \in \mathbf{P}_{c,i}^*} L_{\mathrm{m}}(\mathbf{x}_i^c, \mathbf{x}_j^c; f).$$

  - Negative set

$$L_{\mathrm{N}}(\mathbf{x}_i^c; f) = \sum_{\mathbf{x}_j^k \in |\mathbf{N}_{c,i}^*|} \frac{w_{ij}}{\sum_{\mathbf{x}_j^k \in |\mathbf{N}_{c,i}^*|} w_{ij}} L_{\mathrm{m}}(\mathbf{x}_i^c, \mathbf{x}_j^k; f).$$

  - Combination

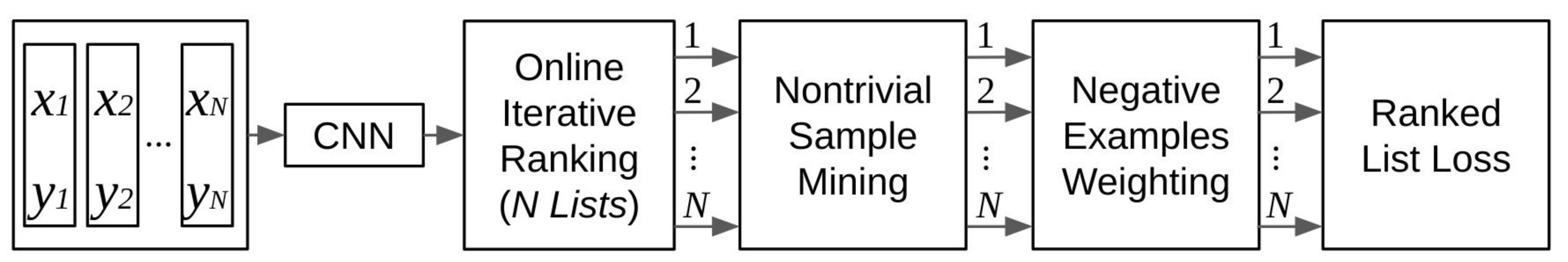
$$L_{\mathrm{RLL}}(\mathbf{x}_i^c; f) = L_{\mathrm{P}}(\mathbf{x}_i^c; f) + \lambda L_{\mathrm{N}}(\mathbf{x}_i^c; f).$$

$$L_{\mathrm{RLL}}(\mathbf{X}; f) = \frac{1}{N} \sum_{\forall c, \forall i} L_{\mathrm{RLL}}(\mathbf{x}_i^c; f).$$

QUEEN'S UNIVERSITY BELFAST anyvision.

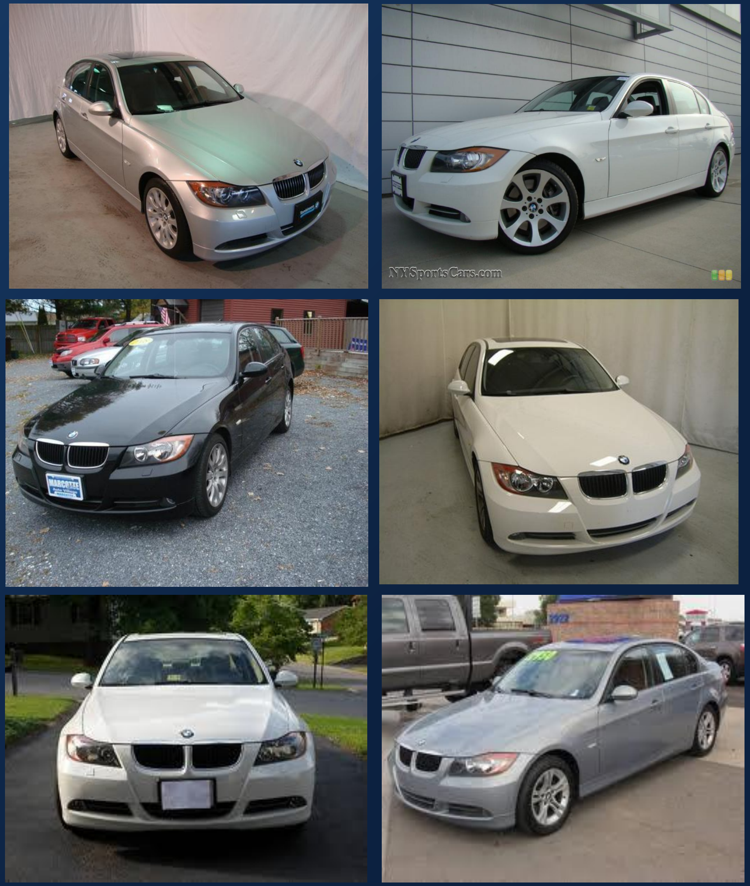
Methodology

- Overall Framework



- Computational Complexity: $O(N^2)$

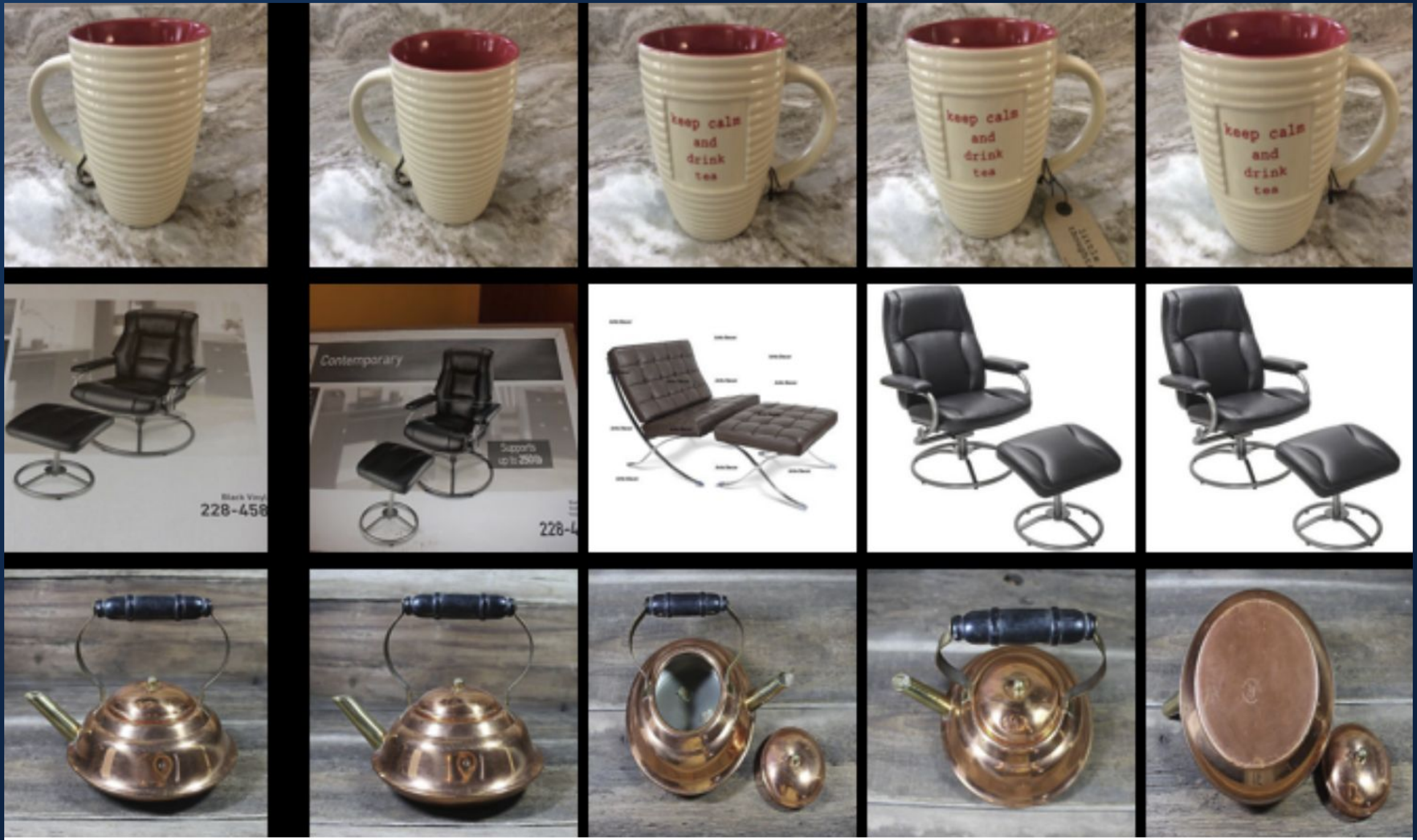
Experiments: Fine-grained Image Retrieval and Clustering Tasks



CARS196 dataset

CUB-200-2011 dataset

Experiments: Fine-grained Image Retrieval and Clustering Tasks



Query The top 4 images in the ranked list of each query

Experiments: Fine-grained Image Retrieval and Clustering Tasks

Evaluation on Fine-grained Image Retrieval and Clustering Tasks

- Results on CARS196 and CUB-200-2011

	CARS196					CUB-200-2011				
	R@1	R@2	R@4	R@8	NMI	R@1	R@2	R@4	R@8	NMI
Triplet Semihard [25]	51.5	63.8	73.5	82.4	53.4	42.6	55.0	66.4	77.2	55.4
Lifted Struct [21]	53.0	65.7	76.0	84.3	56.9	43.6	56.6	68.6	79.6	56.5
N-pair-mc [29]	53.9	66.8	77.8	86.4	57.8	45.4	58.4	69.5	79.5	57.2
Struct Clust [30]	58.1	70.6	80.3	87.8	59.0	48.2	61.4	71.8	81.9	59.2
Spectral Clust [17]	73.1	82.2	89.0	93.0	64.3	53.2	66.1	76.7	85.3	59.2
Proxy NCA [20]	73.2	82.4	86.4	88.7	64.9	49.2	61.9	67.9	72.4	59.5
RLL-H	**74.0**	**83.6**	**90.1**	**94.1**	**65.4**	**57.4**	**69.7**	**79.2**	**86.9**	**63.6**
RLL-(L,M,H)	**82.1**	**89.3**	**93.7**	**96.7**	**71.8**	**61.3**	**72.7**	**82.7**	**89.4**	**66.1**

QUEEN'S UNIVERSITY BELFAST anyvision.

Evaluation on Fine-grained Image Retrieval and Clustering Tasks

- Results on SOP

	SOP			
	R@1	R@10	R@100	NMI
Triplet Semihard [25]	66.7	82.4	91.9	89.5
Lifted Struct [21]	62.5	80.8	91.9	88.7
N-pair-mc [29]	66.4	83.2	93.0	89.4
Struct Clust [30]	67.0	83.7	93.2	89.5
Spectral Clust [17]	67.6	83.7	93.3	89.4
Proxy NCA [20]	73.7	–	–	**90.6**
RLL-H	**76.1**	**89.1**	**95.4**	89.7
RLL-(L,M,H)	**79.8**	**91.3**	**96.3**	90.4

Ablation Studies

- Mining Non-trivial Examples

Table 3: The impact of α on the distance distribution of negative examples. SOP is used. Recall@K (%) results are reported. In all experiments, $m = 0.4, T = 10$.

$m = 0.4, T = 10$	R@1	R@10	R@100
$\alpha = 1.4$	76.2	89.4	95.6
$\alpha = 1.2$	**79.8**	**91.3**	**96.3**
$\alpha = 1.0$	78.7	90.5	95.9

Ablation Studies

- Mining Non-trivial Examples

Table 4: The impact of the distance margin m between negative and positive examples. The Recall@K (%) results on SOP are shown with $\alpha = 1.2, T = 10$ in all experiments.

$\alpha = 1.2, T = 10$	R@1	R@10	R@100
$m = 0$	76.1	89.8	95.7
$m = 0.2$	79.0	91.2	**96.3**
$m = 0.4$	**79.8**	**91.3**	**96.3**
$m = 0.6$	79.2	90.6	96.0
$m = 1.2$	79.1	90.5	95.8

Ablation Studies

- Weighting Negative Examples

Table 5: The results of different T on SOP in terms of Recall@K (%). We fix $m = 0.4, \alpha = 1.2$ in all experiments.

$m = 0.4, \alpha = 1.2$	R@1	R@10	R@100
$T = 0$	78.8	90.7	96.1
$T = 5$	79.1	91.0	96.2
$T = 10$	**79.8**	**91.3**	**96.3**
$T = 15$	79.3	90.9	96.0
$T = 20$	78.6	90.5	95.7

Ablation Studies

- Single-level versus Multilevel Embeddings

Table 6: Single-level embeddings versus multilevel embedding on SOP in terms of Recall@K (%). L, M and H represent the low-level, mid-level and high-level embedding respectively. (L,M,H) means the concatenation of low-level, mid-level and high-level embeddings.

Embedding	R@1	R@10	R@100
L	76.1	88.8	94.9
M	76.9	89.6	95.5
H	76.1	89.1	95.4
(L,M,H)	**79.8**	**91.3**	**96.3**

Ablation Studies

- The Impact of Batch Size

Table 7: The results of different batch size on SOP.

Batch size	R@1	R@10	R@100
$120 = 40 \times 3$	79.2	90.9	96.2
$150 = 50 \times 3$	79.5	91.1	96.2
$165 = 55 \times 3$	79.7	91.2	**96.3**
$180 = 60 \times 3$	**79.8**	**91.3**	**96.3**
$195 = 65 \times 3$	**79.8**	**91.3**	**96.3**

Summary

- Key motivations of RLL

Exploiting all non-trivial training examples

Hypersphere Regularisation:
Non-excessively Reduce Intraclass Variance

Imbalanced number of
positive and negative examples

- Simple, Intuitive & Effective.

Ranked List Loss for Deep Metric Learning

Many Thanks !

Comments Are Appreciated !

Questions Are Welcome !