Collaborative Representation for Deep Meta Metric Learning

Min Zhu
College of Oceanography and Space Informatics, China University of Petroleum (East China)
Qingdao, Shandong, China

Weifeng Liu
College of Control Science and Engineering, China University of Petroleum (East China)
Qingdao, Shandong, China

Kai Zhang
School of Petroleum Engineering, China University of Petroleum (East China)
Qingdao, Shandong, China

Ye Li
Qilu University of Technology (Shandong Academy of Sciences)
Jinan, Shandong, China

Peng Liu
Shandong Kexun Information Technology Co., Ltd
Qingdao, Shandong, China

Baodi Liu*
College of Control Science and Engineering, China University of Petroleum (East China)
Qingdao, Shandong, China

ABSTRACT

Most metric learning methods utilize all training data to construct a single metric, and it is usually over-fitting on the "salient" feature. To overcome this issue, we propose a deep meta metric learning method based on collaborative representation. We construct multiple episodes from the original training data to train a general metric, where each episode consists of a query set and a support set. Then, we introduce a collaborative representation method, which fits the query sample with the support samples per class. We predict the query sample's label via the optimal fitness among the query sample and the support samples in each specific class. Besides, we adopt a hard mining strategy to learn a more discriminative metric according to increasing the training tasks' difficulty. Experiments verify that our method achieves state-of-the-art results on three re-ID benchmark datasets.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**; *Supervised learning by classification*; Image representations.

KEYWORDS

Metric learning; meta learning; collaborative representation; hard mining

ACM Reference Format:
Min Zhu, Weifeng Liu, Kai Zhang, Ye Li, Peng Liu, and Baodi Liu. 2021. Collaborative Representation for Deep Meta Metric Learning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3460426.3463583

*Corresponding author: thu.liubaodi@gmail.com

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463583

1 INTRODUCTION

Recently, deep metric learning has been widely used in various visual tasks and has achieved remarkable success, including image retrieval, image clustering, face recognition, person and vehicle re-identification. It aims to learn a mapping from the original data space to the embedding space [15, 21], so that the intra-class instances in the embedding space have similar representations, and the inter-class instances have different representations.

Contrast loss [4] and triplet loss [1] are widely used in metric learning. Their main idea is to shorten the distance between intra-class instances and increase the distance between inter-class instances. The input of contrast loss is a sample pair, and the input of triplet loss includes three samples. As an extension of triplet loss, n-pair loss [14] considers multiple negative instances to keep a large margin between the anchor sample and all heterogeneous samples. Recently, Movshovitz-Attias et al. [11] have exploited a proxy-based metric approach. It sets a proxy for each class and computes the distance between the anchor sample and each proxy. This method can significantly reduce computational complexity and improve convergence speed.

The method mentioned above regards easy samples and hard samples as equal. They neglect massive useful information and fail to learn efficiently. And easy samples produce gradients that are approximately zero [24]. Therefore, more and more people begin to pay attention to the mining of hard samples in recent years. Hermans et al. [6] proposed a Triplet loss with batch hard mining (TriHardloss). Given an anchor sample, this method aims to select the hardest negative sample and the hardest positive sample to construct a triplet. Its performance is better than the traditional triplet loss. Subsequently, Qiqi Xiao et al. proposed the MSML method employing the hardest negative sample pair and the hardest positive sample pair to calculate the loss [20]. Recently, lifted structured loss is proposed to learn a metric space by using all negative and positive sample pairs in the training set [15].

However, these methods use all training samples to learn a metric space, which usually leads to over-fitting on "salient" features. Recently, Chen et al. proposed the DMML method to train the distance metric, including hard mining distance and center support distance [2], in a meta manner. It can alleviate the over-fitting effectively by sampling multiple episodes from the original training data to learn a general metric. But, DMML also suffers two problems.

First, the intra-class similarity relationships are not considered. Second, only the hardest samples are considered, while the information in other samples that may be helpful to training is ignored.

In this paper, we propose a new deep meta-metric learning approach to tackle the issues above by introducing collaborative representation and a hard mining strategy. First, we construct multiple episodes like the DMML. A query set and a support set are contained in each episode. Second, the collaborative representation is utilized for training the meta-metric space. It fits the query sample with the support samples for each class, and we can predict the query sample's label via the optimal fitness among the query sample and the support samples in each specific class. Finally, we adopt a hard mining strategy to select the top M hard samples containing more potential information in each class to optimize the metric further. In the experiments, we verify the outstanding performance of our method on the Market-1501, DukeMTMC-reID, and VeRi-776 datasets. The main contribution of our work is summarized as:

1) We propose a deep meta metric learning approach by introducing the collaborative representation. Our approach trains a metric in a meta way and can fit the query sample with the support samples per class. It expresses intra-class similarity well and avoids the over-fitting phenomenon effectively.
2) We combine a hard mining strategy to improve the performance further.
3) Experiments show that our approach has achieved state-of-the-art performance and has powerful generalization and discrimination.

2 METHODOLOGY

2.1 Problem Formulation

Unlike conventional metric learning methods, we formulate metric space in a meta manner [13, 22], and the training data is a collection of different but related subtasks. Specifically, we randomly sample K classes from the original dataset X to construct an episode $D \subseteq X$. There are P instances in each class. Then, the training data D is randomly divided into a query set Q and a support set S, where $S \cup Q = D$. Finally, we can obtain a general metric applicable to all subtasks. The overall objective of our method is expressed as:

$$\Theta = arg\min_{\Theta} \sum_{t \in T_{train}} \sum_{(S,Q) \in D^t} L^t\left(S, Q; \Theta\right) \tag{1}$$

where T_{train} is the number of episodes, and D^t represents the training data to episode t.

2.2 Collaborative Representation Method

The collaborative representation can be regarded as a function-fitting training process, which fits the query sample with the support samples for each class. Then, the query sample is assigned to the class with the optimal fitness. It fully considers the similarities between the intra-class and the differences between the inter-class.

Suppose there is a training set $X = [X^1, X^2, ..., X^C] \in \mathbb{R}^{d \times N}$, where $X^c \in \mathbb{R}^{d \times N_c}$ is the feature representation from class c and d is the feature dimension per instance. $N = \sum_{c=1}^{C} N_c$ represents the total number of instances over C classes, and N_c represents the number of instances from the c_{th} class. In this work, we set $N_c > d$, which is over-determined.

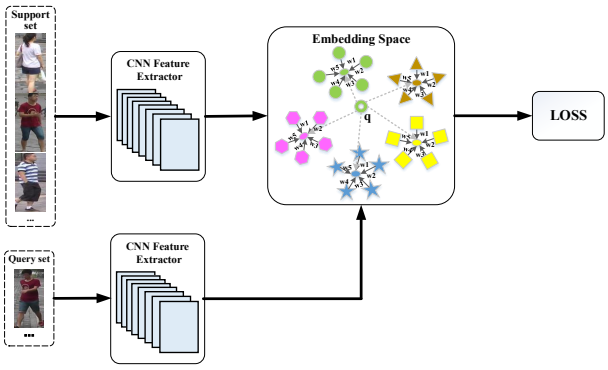

Figure 1: The structure of our method. For each episode, training samples consist of a query set and a support set. We obtain a fitting feature using the collaborative representation and compute the similarity between the query instance and fitting feature per class.

Given a query sample $y \in \mathbb{R}^{d \times 1}$, this method aims to solve the following minimization problem:

$$\hat{W} = arg\min_{W} \left\|y - X^c W^c\right\|_2^2 + \beta \left\|W^c\right\|_2^2 \tag{2}$$

where β is the regularization parameter, and $W^c \in \mathbb{R}^{N_c \times 1}$ represents the weight matrix corresponding to class c.

Then, a closed-form solution is calculated by the least square method:

$$\hat{W}^c = \left(X^{cT} X^c + \beta I\right)^{-1} X^{cT} y \tag{3}$$

The residual error between the query sample y and each class can be calculated as:

$$r_c\left(y\right) = \left\|y - X^c \hat{W}^c\right\|_2^2 \tag{4}$$

where, $X^c \hat{W}^c$ represents the weighted average of each class. The final class of the query sample y is obtained by:

$$C\left(y\right) = arg\min_{c} r_c\left(y\right) \tag{5}$$

2.3 Overall algorithm

Fig. 1 describes the overall process of our approach. We sample multiple episodes from the given datasets and learn a metric in each episode. Specifically, we learn a weight for each support sample during training and then use the optimal weighted average of the support sample to represent each class. This is achieved by using a collaborative representation learning. The method applies all instances in the support set and considers the relative hardness of the samples. That is, hard samples occupy a more massive weight than easy samples.

However, easy samples contain less useful information and have little effect on promoting performance. Moreover, considering easy samples increases the computational complexity and reduces hard samples' weight, making it impossible to mine rich data information and optimize the metric space well. To learn a better discriminative

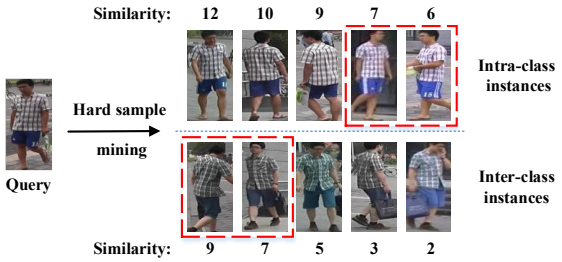

Figure 2: The illustration of hard mining strategy. The red dotted line represents the hard samples we mined.

feature and use sample information more efficiently, we introduce a hard mining strategy. As shown in Fig. 2, we mine the top M hard samples that contain abundant information in each class. The M samples from the same class as the query have lower similarities, and the M samples from a different class have higher similarities. Then, we employ the M hard support samples as the input to learn a better embedding space. Algorithm 1 illustrates the step of our approach in detail.

Algorithm 1 Collaborative representation for deep meta metric learning.

Require: Training data X, episode size T_{train}, class number K, support instances and query instances number per class.
Ensure: Embedding space parameters θ;
1: Initialize parameter θ;
2: **for** $t = 1, 2, \cdots, T_{train}$ **do**
3: Randomly sample K classes from X;
4: Construct a query set and support set for each class;
5: Mine hard support samples per class;
6: Compute the weight matrix of hard samples for each class following (3);
7: Compute the optimal weighted average of support samples per class using $X^c\hat{W}^c$;
8: Compute the residual error between the query sample and the optimal weighted representation of each class following (4);
9: Optimize θ_t by taking the θ_{t-1} as the initial guess.
10: **end for**
11: **return** θ

3 EXPERIMENTS

3.1 Datasets and settings

Datasets. Market-1501 [23], DukeMTMC-reID [12, 16], and VeRi-776 [8, 9] are used to evaluate the proposed method. VeRi-776 is a vehicle re-identification dataset, and the rest are person re-identification datasets. All datasets contain a test set and a training set. Table 1 displays the detailed statistics of these datasets.

Training schemes. We use ResNet-50 [5] as the backbone network initialized by the pre-training weights of ImageNet [3]. We resize all images to 256×128 on person datasets and all images to 224×224 on vehicle datasets. Besides, we employ the following data augmentation methods: random erasing, normalization, and random horizontal flip. The random sampling strategy of each episode is set to K-way-P-shot. Here, we recommend the following Settings: $K = 32$, $P = 5$. When the hard sample mining method is introduced, we set $K = 32$, $P = 7$. Given a query, we mine the top two hard support samples from each class. We set epoch to 1, 200 and L2 weight decay to 10^{-5}. Adam is used to optimizing the network. The initial learning rate is set to 0.0002. It begins to decline exponentially after 600 epochs and finally decreased to 10^{-6}. We set regularization parameter β in Eq. (4) to 2.0 and 4.0 respectively on person and vehicle datasets.

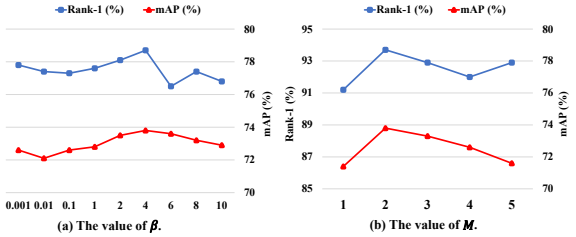

Figure 3: Influence of parameter β and M for Rank-1 accuracy and mAP on the VeRi-776 dataset.

Evaluation protocol. This paper employs the cumulative matching characteristic (CMC) curve and mean average precision (mAP) to evaluate the proposed method's performance. Specifically, we present the CMC accuracy of the proposed method on rank-1, rank-5, and rank-10. Besides, we do not apply the re-ranking [25] method and only adopt the single query mode.

3.2 Results and analysis

Comparison with Baseline methods. We compare the proposed method with several baseline methods on three re-identification datasets. Experimental results are presented in Table 2. It can be seen that our approach has great superiority compared with the baseline method. Our approach achieves 93.1%, 85.3%, 93.7% respectively at rank-1 on Market-1501, DukeMTMC-reID and VeRi-776 datasets. And we achieve 82.4%, 71.3%, 73.8% respectively at mAP on Market-1501, DukeMTMC-reID and VeRi-776 datasets. In particular, the improvement on mAP is larger than on rank-1.

Comparison with State-of-the-art methods. The statistical comparison with state-of-the-art methods on Market-1501 and DukeMTMC-reID is shown in Table 3. * denotes attention-related methods, and † denotes meta-learning methods. The results illustrate that our approach obtains consistent improvements in the two datasets and is better than most state-of-the-art methods.

The importance of Collaborative representation. Table 4 illustrates the comparisons between our approach and the DMML. Our approach utilizes the collaborative representation (CR) to obtain the optimal weighted average of support samples per class. DMML applies hard mining distance and center support distance to train the network. Specifically, center support distance computes the average of support samples. Hard mining distance aims to seek the nearest sample from the different classes and the farthest sample from the same class with the query sample. Our method has a noticeable improvement at rank-1 and mAP in comparison with DMML on three Re-ID datasets. The performance improves 0.7% and 1.4% respectively at rank-1 and mAP on Market-1501 when

Table 1: Statistics of Re-ID datasets.

Dataset	Identities	Cameras	Train set		Test set		
			identities	images	identities	query	gallery
Market-1501	1501	6	751	12936	750	3368	19732
DukeMTMC-reID	1404	8	702	16552	702	2228	17661
Veri-776	776	20	576	37778	200	1678	11579

Table 2: Comparison with baseline methods on three re-ID benchmark datasets.

Methods	Market-1501				DukeMTMC				VeRi-776		
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	mAP
Contrastive	75.8	88.6	92.4	58.9	68.1	81.4	85.1	49.5	67.4	85.0	49.8
Triplet	89.6	96.2	97.6	76.2	80.7	90.7	93.1	65.4	90.0	95.2	68.1
N-pair	89.4	96.1	97.6	77.4	82.0	91.9	94.4	68.3	88.6	95.1	65.1
Softmax	86.7	94.5	96.6	70.2	77.0	87.7	91.7	59.6	87.4	94.6	57.8
Softmax+Center Loss	91.2	96.5	97.9	77.6	82.3	91.7	93.6	66.3	90.8	95.6	66.0
Proxy-NCA	88.0	95.4	97.1	71.0	77.9	88.2	91.6	58.1	86.7	93.3	56.4
LiftedStruct	90.5	96.8	98.0	78.4	82.6	91.2	93.8	68.0	90.8	96.1	69.3
DMML	92.4	97.3	98.3	81.0	84.3	92.6	94.6	70.2	91.2	96.3	70.1
Ours	**93.1**	**97.6**	**98.7**	**82.4**	**85.3**	**93.3**	**95.1**	**71.3**	**93.7**	**97.3**	**73.8**

Table 3: Comparison with the state-of-the-art methods.

Method	Issue	Market-1501		DukeMTMC	
		R-1	mAP	R-1	mAP
SVDNet [18]	ICCV'17	82.3	62.1	76.7	56.8
PCB [19]	ECCV'18	92.3	77.4	81.8	66.1
*HA-CNN [7]	CVPR'18	91.2	75.7	80.5	63.8
*VPM [17]	CVPR'19	93.0	80.8	83.6	72.6
PGFA [10]	ICCV'19	91.2	76.8	82.6	65.5
†DMML [2]	ICCV'19	92.4	81.0	84.3	70.2
M^3P-RID [26]	CVPR'20	95.4	82.6	84.7	68.5
Ours	-	**93.1**	**82.4**	**85.3**	**71.3**

Table 4: Comparison with DMML on Market-1501, DukeMTMC-reID and VeRi-776 datasets.

Dataset	Method	Ours		DMML	
		R-1	mAP	R-1	mAP
Market1501	CR/Center	89.5	74.7	87.1	70.3
	+ Hard mining	**93.1**	**82.4**	92.4	81.0
DukeMTMC	CR/Center	81.1	64.1	78.5	61.6
	+ Hard mining	**85.3**	**71.3**	84.3	70.2
VeRi776	CR/Center	89.9	65.9	89.1	65.0
	+ Hard mining	**93.7**	**73.8**	91.2	70.1

considering hard sample mining. Our method achieves consistent improvements on other datasets. The results demonstrate the importance and superiority of collaborative representation.

3.3 Parameter Analysis.

We discuss the impact of hyper-parameters β and M on the veri-776 dataset in this section. Fig.3 shows the performance of our method under different settings.

Regularization parameter β. Fig.3(a) reports the performance with different values of β. As β increases, rank-1 accuracy and mAP present a trend of rising at first and then declining. It can be seen that the performance is optimal when $\beta = 4.0$. Therefore, $\beta = 4.0$ is adopted in our experiments.

Number of hard mining instances M. The number of hard mining instances plays a crucial role in improving the rank-1 accuracy and mAP. As shown in Fig.2(b), we set M from 1 to 5 to evaluate our proposed approach. $M = 1$ denotes that only the hardest instance is mined to optimize the metric. Its accuracy is low due to ignoring rich information in other samples. It can be seen that the performance reached the peak point when $M = 2$. As M continues to increase, the performance begins to decline due to introducing some easy samples with less valid information. Therefore, we set $M = 2$ in our experiments.

4 CONCLUSION

In this paper, we propose a deep meta-metric learning method by introducing collaborative representation. It can effectively avoid the over-fitting phenomenon and learn a better feature space where similar samples have similar representations. We sample multiple subtasks from the training data and randomly divide them into a support set and query set at first. Then we introduce a collaborative representation method to obtain the optimal weighted average of the support samples and optimize the meta metric further. Finally, we consider hard sample mining to promote the speed of model convergence. We demonstrate the powerful generalization and discrimination of the proposed method on person and vehicle re-identification datasets.

ACKNOWLEDGMENT

The paper was supported by the Natural Science Foundation of Shandong Province, China (Grant No. ZR2019MF073), the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) (Grant No. 20CX05001A), the Major Scientific and Technological Projects of CNPC (No. ZD2019-183-008), and the Creative Research Team of Young Scholars at Universities in Shandong Province (No.2019KJN019).

REFERENCES

[1] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large Scale Online Learning of Image Similarity through Ranking. *Journal of Machine Learning Research*, 1109–1135.

[2] Guangyi Chen, Tianren Zhang, Jiwen Lu, and Jie Zhou. 2019. Deep Meta Metric Learning. In *ICCV*. 9546–9555.

[3] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. 248–255.

[4] R. Hadsell, S. Chopra, and Y. Lecun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*. 1735–1742.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737* (2017).

[7] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious Attention Network for Person Re-Identification. In *CVPR*. 2285–2294.

[8] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2016. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *ECCV*. 869–884.

[9] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2017. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia* 20, 3 (2017), 645–658.

[10] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In *ICCV*. 542–551.

[11] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. 2017. No Fuss Distance Metric Learning using Proxies. In *ICCV*. 360–368.

[12] Ergys Ristani, Francesco Solera, Roger S Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *ECCV*. 17–35.

[13] Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical Networks for Few-shot Learning. *arXiv preprint arXiv:1703.05175* (2017).

[14] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class N-pair loss objective. In *NeurIPS*. 1857–1865.

[15] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *CVPR*. 4004–4012.

[16] Wenchen Sun, Fangai Liu, and Weizhi Xu. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *ICCV*. 3774–3782.

[17] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 2019. Perceive Where to Focus: Learning Visibility-aware Part-level Features for Partial Person Re-identification. In *CVPR*. 393–402.

[18] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. 2017. SVDNet for Pedestrian Retrieval. In *ICCV*. 3800–3808.

[19] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*. 480–496.

[20] Qiqi Xiao, Hao Luo, and Chi Zhang. 2017. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. *arXiv preprint arXiv:1710.00478* (2017).

[21] Baosheng Yu and Dacheng Tao. 2019. Deep Metric Learning With Tuplet Margin Loss. In *ICCV*. 6489–6498.

[22] Fahong Zhang, Qi Wang, and Xuelong Li. 2020. Deep Meta-Relation Network for Visual Few-Shot Learning. In *ICASSP 2020*. 1509–1513.

[23] Liang Zheng, Liyue Shen, Shengjin Tian, Lu abibnd Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*. 1116–1124.

[24] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. 2019. Hardness-Aware Deep Metric Learning. In *CVPR*. 1–1.

[25] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking Person Re-identification with k-reciprocal Encoding. In *CVPR*. 3652–3661.

[26] Jiahuan Zhou, Bing Su, and Ying Wu. 2020. Online Joint Multi-Metric Adaptation From Frequent Sharing-Subset Mining for Person Re-Identification. In *CVPR*. 2906–2915.