DEEP FEATURE EMBEDDING LEARNING FOR PERSON RE-IDENTIFICATION USING LIFTED STRUCTURED LOSS

Zhangping He[†]*, Zhendong Zhang*[†] *and Cheolkon Jung*

School of Electronic Engineering, Xidian University, Xian, Shaanxi 710071, China
zhengzk@xidian.edu.cn

ABSTRACT

In this paper, we propose deep feature embedding learning for person re-identification (re-id) using lifted structured loss. Although triplet loss has been commonly used in deep neural networks for person re-id, the triplet loss-based framework is not effective in fully using the batch information. Thus, it needs to choose hard negative samples manually that is very time-consuming. To address this problem, we adopt lifted structured loss for deep neural networks that makes the network learn better feature embedding by minimizing intra-class variation and maximizing inter-class variation. Extensive experiments on CUHK03, CUHK01 and VIPeR datasets demonstrate the superior performance of the proposed method over state-of-the-arts in terms of the cumulative match curve (CMC) metric.

Index Terms— Person re-identification, convolutional neural networks, deep learning, lifted structured loss, triplet loss

1. INTRODUCTION

Person re-id or person retrieval, which matches pedestrians from different video cameras, has wide application in video surveillance. It is a challenging task because of large variations in viewpoint and lighting across different views. Most existing methods primarily focus on either feature extraction or similarity measurement. Feature extraction is to find features that are robust to challenging factors as well as discriminative to different identities. Unfortunately, it is still extremely hard to design a feature that is distinct, reliable and invariant to severe variations and misalignment across disjoint views. Similarity measurement aims to learn a optimal metric under which instances belonging to the same person are closer than different persons. These approaches typically extract hand-crafted features from the training dataset, and subsequently learn the metrics. However, if feature representation is not reliable, some useful information would be lost in the first step, and it cannot be expected that the learned metric in the second step would have desirable performance. Thus, it would be a good choice to jointly learn feature representation and metric. Recently, thanks to the available of large person re-id datasets such as CUHK03 [1] and Market1501 [2], and the great success of deep learning approaches in various computer vision tasks, a lot of deep learning-based person re-id approaches have been proposed [3, 4], which have made great progress in performance. There are two common frameworks in deep learning-based re-id. One of them is to use a part of the network as a feature extractor and to measure the similarity between two images with metric learning such as image pair verification loss [5]. These models are trained in cross-image representation, resulting in expensive computational cost in a real test. This is because each probe needs to go through the network paired with every gallery image. Other deep learning-based methods directly train a desired deep feature embedding which minimizes intra-class distance and maximizes inter-class distance. Among them, deep learning approaches with a contrastive or triplet loss becomes a popular framework for person re-id and demonstrates superior performance [6, 7, 8]. The contrastive loss or triplet loss makes it possible to perform end-to-end learning between the input images and the desired embedding space such that the same person are mapped into nearby points while different people are mapped apart from each other. However, the results are often unsatisfactory if we naively apply it to the person re-id problem. This is because the possible number of triplets grows exponentially as the number of dataset increases. Most of them are redundant, which makes training quickly stagnate. Therefore, some researchers take a hard triplet mining strategy [6, 7]. However, mining such hard triplets is time-consuming, and selecting too hard triplets often makes the training procedure unstable. Thus, Shi *et al.* [9] have proposed a moderate method for positive sample mining to select positive samples that are between the hardest positive samples and the hardest negative samples, which shows better and more stable training results.

In this paper, we propose deep feature embedding learning for person re-id using lifted structured loss. We apply the lifted structured loss [10] to person re-id that fully uses the information of each batch, and thus both hard sample mining and moderate sample mining are not needed. We verify that

[†] Authors contributed equally. This work was supported by the National Natural Science Foundation of China (No. 61271298) and the International S&T Cooperation Program of China (No. 2014DFG12780).

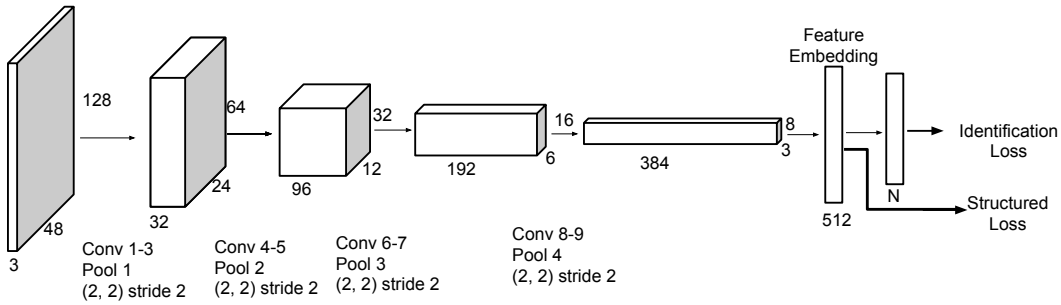

Fig. 1. Illustration of the proposed network architecture.

the lifted structured loss is superior to naive contrastive loss and triplet loss in both accuracy and training speed. Moreover, we combine the lifted structure loss and identification loss that are complementary [11, 12, 13]. Experimental results demonstrate that the proposed method achieves better performance than the state-of-the-art ones for person re-id. Fig. 1 illustrates the entire framework of the proposed person re-id based on lifted structure loss. Compared with the existing methods, the main contributions of this paper are summarized as follows:

- We adopt the lifted structured loss for person re-id and verify its superiority over contrastive and triplet losses.
- We combine the lifted structured loss with the identification loss to consider both relative information of sample pairs (positive or negative) and true identity information.

2. PROPOSED METHOD

Triplet Loss [14] was first introduced by F. Schroff for face recognition and clustering, which is trained on the triplet data $\left\{\left(x_a^i, x_p^i, x_n^i\right)\right\}$ where $\left(x_a^i, x_p^i\right)$ is from the same class, while the term of $\left(x_a^i, x_n^i\right)$ is from different classes. Intuitively, the metric embedding learning encourages to learn a function f which maps semantically similar points to the data manifold onto metrically close points, and maps semantically different points from the data manifold onto metrically distant points. The loss function is defined as follows:

$$L_{triplet} = \frac{1}{2m}\sum_{i=1}^{m}\left[D_{ia,ip}^2 - D_{ia,in}^2 + \alpha\right]_+ \qquad (1)$$

where $D_{ia,ip}^2 = \left\|f\left(x_a^i\right) - f\left(x_p^i\right)\right\|_2^2$, m is the batch size, and $\alpha > 0$ controls the margin of triplet loss. The $[\cdot]_+$ operation indicates the hinge function $\max(0, \cdot)$. In this work, we normalize the learned features f so that the range of D^2 is [0,4]. It can be observed that traditional triplet loss is not able to make full of the training batch. If the batch size is m, then the number of triplets is $m/3$. To solve this problem, the lifted structured loss has been proposed by Song *et al.*.

Lifted Structured Loss [10] boosts the vector of pairwise distances within the batch to the matrix of pairwise distances to take full advantage of the training batch. The loss function is defined as follows:

$$\begin{aligned} L_{i,j} &= \max\left(\max_{(i,k)\in\hat{N}} \alpha - D_{i,k}, \max_{(j,l)\in\hat{N}} \alpha - D_{j,l}\right) + D_{i,j} \\ L_{lifted} &= \frac{1}{2|\hat{P}|}\sum_{(i,j)\in\hat{P}} \max\left(0, L_{i,j}\right)^2 \end{aligned} \qquad (2)$$

where $\hat{P}$ is the set of positive pairs and $\hat{N}$ is the set of negative pairs in the training set. The lifted structured loss makes full use of the batch by transforming a training batch of samples into a fully connected dense matrix of pairwise distances. However, this loss causes two computational challenges: (1) It is non-smooth, and (2) both evaluating it and computing its sub-gradient require to traverse all pairs of examples several times. Thus, a smooth upper bound on the function is used for the lifted structured loss is defined as follows:

$$\begin{aligned} \widetilde{L}_{i,j} &= \log\left(\sum_{(i,k)\in\hat{N}} e^{\alpha - D_{i,k}} + \sum_{(j,l)\in\hat{N}} e^{\alpha - D_{j,l}}\right) + D_{i,j} \\ \widetilde{L}_{lifted} &= \frac{1}{2|\hat{P}|}\sum_{(i,j)\in\hat{P}} \max\left(0, L_{i,j}\right)^2 \end{aligned} \qquad (3)$$

Improvement of Lifted Structured Loss: It can be observed that the number of summation terms is uncertain in the first formula in Eq. (3) since the number of negative samples relative to the positive pairs is variable. Thus, it is not able to balance between $\log$ term and $D_{i,j}$. To tackle this problem, we propose to calculate the mean of $\log$ term so that the range of log term in $\widetilde{L}_{i,j}$ keeps within $[\alpha - 4, \alpha]$ by preventing the variation with the number of negative pairs. Furthermore, replacing $D_{i,j}$ with $D_{i,j}^2$ makes training converge easily and achieves better results. Consequently, our loss function which

is defined as follows:

$$L_{i,j} = \log\left(\frac{1}{|\hat{T}_{i,j}|}\left(\sum_{(i,k)\in\hat{N}} e^{\alpha - D^2_{i,k}} + \sum_{(j,l)\in\hat{N}} e^{\alpha - D^2_{j,l}}\right)\right) + D^2_{i,j}$$
$$L_{struct} = \frac{1}{2|\hat{P}|}\sum_{(i,j)\in\hat{P}} \max\left(0, L_{i,j}\right) \tag{4}$$

where $|\hat{T}_{i,j}|$ is the number of negative samples corresponding to positive pairs $\{i, j\}$. Triplet loss and lifted structured loss only constrain the relative distance between samples. While the identification label contains true identity information, which is complementary with the structured loss. Therefore, we define the loss function of the proposed network architecture based on the combination of them as follows:

$$\begin{aligned} q_i &= softmax\left(W_i^T f\left(x\right)\right) \\ L_{id} &= \sum_i -p_i \log q_i \\ L &= L_{struct} + \lambda L_{id} \end{aligned} \tag{5}$$

where W_i is the i-th row of the parameter matrix of the last classification layer, p_i is the i-th value of identification label, and λ is the parameter to balance two losses.

Network Architecture: Our CNN model is similar to [4] for fair comparison. It is a single branch CNN which consists of nine convolutional (Conv) layers, four max pooling (Pool) layers, one fully connected (FC) layer, and a softmax classification layer as illustrated in Fig. 1. All Conv layers use 3×3 filters with stride 1 and zero padding. All max pooling layers have 2×2 filters with stride 2. Batch normalization is applied after each Conv layer or FC layer to speed up training. Then, LReLU is used after these layers as the non-linear activation function. After the first FC layer, we obtain a 512 dimensional vector which is the feature embedding constrained by the structured loss. Finally, we add a softmax classification layer with N nodes, i.e. the number of the identities.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

We perform our experiments on three publicly available datasets: CUHK03 [1], CUHK01 [15] and VIPeR [16]. For CUHK03 [1], we simply train the model on its training dataset using stochastic gradient descent with mini-batches. Since the size of CUHK01 [15] and VIPeR [16] is small, we adopt a deep transfer learning method similar to [17]. We first pre-train the model on large person re-id datasets that consists of CUHK03 [1] and Market1501 [2], then fine-tune it on the corresponding training set. Note that a two-stepped fine-tuning strategy from [17] is used in this work to conduct more effective transfer learning.

Data Preparation: For all datasets, we resize all training images to 128×48. Similar to [5], we sample 3 images around an image center with small translation as well as augment the data with images reflected on a vertical mirror. Finally, we get 5 images with the size of 128×48 from the original training image. All test images are resized to 128×48. The mean of training data is subtracted by all images.

Evaluation Protocol: We adopt the widely used cumulative match curve (CMC) metric [1] for quantitative evaluations. Our evaluation is in single-shot. Single-shot setting assumes that there exists only one image instance per person in each camera viewpoint. For CUHK01 [15] and VIPeR [16], we randomly select half persons for training and the remaining half for testing. For CUHK03 [1], we randomly select 1260 persons for training and the remaining 100 persons for testing following the protocol used in [1]. In the testing stage, we input all the testing images to the CNN model to get feature embedding, i.e. the output of FC1, for each of them. Then, we normalize each embedding to an unit vector and computer the CMC by ranking the $L2$ distance between query and gallery features. We set λ to 1.0 and the initial learning rate is 0.001, decayed by 0.1 after 20,000 iterations. We set α in the structured loss to 3.0 and α in contrastive loss and triplet loss to 1.0. Batch size is set to be 64 and the iteration number is 30,000. We conduct two sets of experiments: 1) Evaluating the proposed loss with other losses (*e.g.* contrastive loss and triplet loss); 2) Comparing the proposed method with the state-of-the-art re-id ones.

3.2. Performance Comparison Between Different Loss Functions

To evaluate the performance of different loss functions, we conduct experiments in the same CNN architecture. The only difference is that contrastive loss architecture is a two-branch siamese network and triplet loss architecture is a three-branch siamese network. Fig. 2 shows performance comparison between them on CUHK03. It verifies that the proposed structured loss is superior to triplet loss and contrastive loss. Moreover, the combination of structured loss and identification loss achieves better performance in person re-id.

3.3. Performance Comparison with the State-of-the-Arts

CUHK03: The CUHK03 [1] dataset contains 13,164 pedestrian images, which were taken from 1,360 persons by two surveillance cameras. On average, there are 4.8 images per identity in each view. The dataset provides both manually cropped bounding boxes and automatically cropped ones by a pedestrian detector. We show experimental results on both versions of the data, i.e. Labeled and Detected. From Table 1, it can be seen that the proposed method achieves the best accuracy in rank 1, rank 5 and rank 10 on Labeled CUHK03,

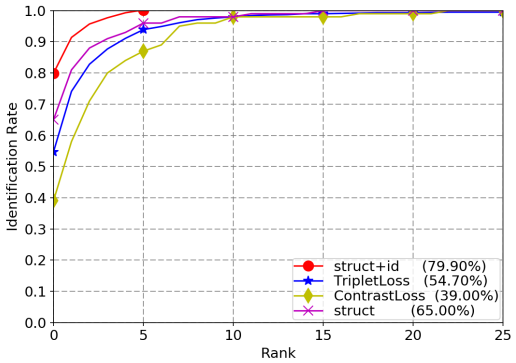

Fig. 2. Comparisons between different loss functions.

Table 1. Accuracy Comparison on CUHK03 (Labeled)

methods	rank1	rank5	rank10
kLFDA[18]	48.20	59.34	66.38
IDLA[5]	54.74	86.50	94.00
NullRe-id[19]	58.90	85.60	92.45
Ensembles[20]	62.10	89.10	94.30
Gated Siamese[3]	68.10	88.10	94.60
NX-Corr M[21]	72.43	95.51	98.40
Proposed	**81.9**	**96.7**	**98.7**

over most latest state-of-the arts. Also, our method outperforms most state-of-art ones in terms of rank 5 and rank 10 on Detected CUHK03 as shown in Table 3, while performs a little worse than [4] in rank1. Note that we do not compare with [17] because they use extra ImageNet data for pre-training.

CUHK01: CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Each person has four images with two from each camera. Table 2 shows person re-id results by the proposed method. Compared with the state-of-the-art methods, the proposed method achieves the best accuracy in rank 1, rank 5 and rank 10.

VIPeR: VIPeR dataset contains two views of 632 persons. It is one of the most challenging dataset for the person re-id task because there are only 316 identities for training with one image per person in each view, giving a total of 316 positive samples. Table 4 compares performance between the proposed method and other ones. Experimental results show that the proposed method as a CNN-based work performs worse than traditional ones such as Least Square Semi-Coupled Dictionary Learning (LSSCDL) [22] and Spatially Constrained Similarity function on Polynomial feature map (SCSP) [23]. This is because deep learning-based methods are not effective in the dataset of small size.

Table 2. Accuracy Comparison on CUHK01

methods	rank1	rank5	rank10
IDLA[5]	47.5	71.6	80.3
NullRe-id[19]	69.1	86.9	91.8
MCP-CNN[7]	53.7	84.3	91.0
NX-Corr M[21]	65.04	89.76	94.4
Proposed	**70.2**	**90.2**	**95.5**

Table 3. Accuracy Comparison on CUHK03 (Detected)

methods	rank1	rank5	rank10
IDLA[5]	45.0	76.0	83.5
NullRe-id[19]	53.70	83.05	93.00
Siamese LSTM[24]	57.3	80.1	88.3
Joint Learning[25]	52.17	85.00	92.00
Gated Siamese[3]	61.8	80.9	88.3
NX-Corr M[21]	72.04	96.00	98.26
Improved Embedding [4]	**82.1**	96.2	98.2
Proposed	79.9	**97.1**	**98.7**

Table 4. Accuracy Comparison on VIPeR

methods	rank1	rank5	rank10
Joint Learning[25]	35.8	-	-
Gated Siamese[3]	37.8	66.9	77.4
Siamese LSTM[24]	42.4	68.7	79.4
Ensembles[20]	45.9	77.5	88.9
MCP-CNN[7]	47.8	74.7	84.8
SCSP[23]	**53.5**	82.6	91.5
NullRe-id[19]	51.2	82.1	90.5
Improved Embedding[4]	50.4	77.6	85.8
LSSCDL[22]	42.7	**84.3**	**91.9**
Proposed	47.3	76.6	88.1

4. CONCLUSIONS

In this paper, we have proposed deep feature embedding learning for person re-id based on lifted structured loss. The proposed person re-id is based on CNN, which combines lifted structured loss and identification loss into the loss function. Our lifted structured loss minimizes the influence of sample distribution on training. In testing stage, we perform feature embeddings on all test images using CNN. Then, we normalize each embedding into a unit vector and compute L_2 distance between all pairs from two camera views, resulting in efficient computation. Experimental results demonstrate that the proposed method achieves better performance than the state-of-art ones on CUHK01 and CUHK03 while performing a little worse than LSSCDL, SCSP and NullRe-id on VIPeR, i.e. a very small dataset.

5. REFERENCES

[1] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[3] Rahul Rama Varior, Mrinal Haloi, and Gang Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 791–808.

[4] Haibo Jin, Xiaobo Wang, Shengcai Liao, and Stan Z Li, "Deep person re-identification with improved embedding," *arXiv preprint arXiv:1705.03332*, 2017.

[5] MJ Jones and TK Marks, "An improved deep learning architecture for person re-identification," 2015.

[6] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[7] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[9] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li, "Embedding deep metric for person re-identification: A study against large variations," in *European Conference on Computer Vision*. Springer, 2016, pp. 732–748.

[10] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4004–4012.

[11] De Cheng, Yihong Gong, Weiwei Shi, and Shizhou Zhang, "Person re-identification by the asymmetric triplet and identification loss function," *Multimedia Tools and Applications*, pp. 1–18, 2017.

[12] Zhedong Zheng, Liang Zheng, and Yi Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, pp. 13, 2017.

[13] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "A multi-task deep network for person re-identification.," in *AAAI*, 2017, vol. 1, p. 3.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[15] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3594–3601.

[16] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*. Citeseer, 2007, vol. 3, pp. 1–7.

[17] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian, "Deep transfer learning for person re-identification," *arXiv preprint arXiv:1611.05244*, 2016.

[18] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[19] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1239–1248.

[20] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1846–1855.

[21] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal, "Deep neural networks with inexact matching for person re-identification," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 2667–2675.

[22] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan, "Sample-specific svm learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1278–1287.

[23] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1268–1277.

[24] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang, "A siamese long short-term memory architecture for human re-identification," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 135–153.

[25] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.