Supplementary Material for "Unsupervised Person Re-identification by Soft Multilabel Learning"

Hong-Xing Yu[1], Wei-Shi Zheng*[1,4],
Ancong Wu[1], Xiaowei Guo[2], Shaogang Gong[3], and Jian-Huang Lai[1]
[1]Sun Yat-sen University, China
[2]YouTu Lab, Tencent
[3]Queen Mary University of London
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
xKoven@gmail.com, wszheng@ieee.org, wuancong@mail2.sysu.edu.cn, scorpioguo@tencent.com,
s.gong@qmul.ac.uk, stsljh@mail.sysu.edu.cn

Abstract

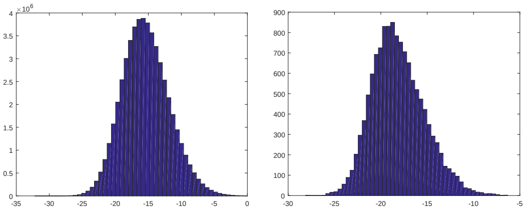
This supplementary material accompanies our main manuscript "Unsupervised Person Re-identifictaion by Soft Multilabel Learning", including the observation of the log-normal distribution of the soft multilabels, the simplified 2-Wasserstein distance, more implementation details and the evaluations on the mining ratio and a hyperparameter β.

1. Observation of the log-normal distribution

As mentioned in Sec. 3.3 in our main manuscript, we empirically observe that the soft multilabel approximately follows a log-normal distribution. Note that we do *not* claim the generality of the log-normal distribution, This empirical observation only serves to facilitate our implementation and computation in computing the distributional distance in the Cross-view soft Multilabel Learning, which does *not* require any specific distribution of the soft multilabel (Please refer to L_{CML} in Eq. (5) in our main manuscript). We show the distribution of the log-soft multilabel in the training set of the Market-1501 [12] dataset in Figure S1(a). From Figure S1(a) we observe that the soft multilabel approximately follows a normal distribution in the log space over all the dimensions. We have also made similar observations across each single dimension. For example, we show the distribution of the first dimension (i.e. the label likelihood w.r.t. the first reference person) in Figure S1(b), which also approximately follows a normal distribution.

2. The simplified 2-Wasserstein distance

As mentioned in Sec. 3.3 in our main manuscript, in this work we adopt the simplified 2-Wasserstein distance [1, 3] which gives a very simple form of the Cross-view consistent soft Multilabel Learning (L_{CML}) in Eq. (6) in our main manuscript.



(a) Over all dimensions (b) The first dimension

Figure S1. The distributions of the log-soft multilabel in the training set of the Market-1501 [12] dataset, using the ResNet-50 backbone. Note that we do not claim the generality of the log-normal distribution. The distribution in (a) includes all dimensions, i.e. the label likelihood w.r.t. all 4101 reference persons in the MSMT17 [11] dataset. The distribution in (b) includes only the first dimension, i.e. the label likelihood w.r.t. the first ID in MSMT17.

Given two normal distributions $X = \mathcal{N}(\mu_x, C_x)$ and $Y = \mathcal{N}(\mu_y, C_y)$ where μ_x is the mean vector of X and C_x is the covariance matrix of X. The 2-Wasserstein distance between X and Y is defined by [1]:

$$W_2(X,Y)^2 = ||\mu_x - \mu_y||_2^2 + \text{trace}(C_x + C_y - 2(C_x^{\frac{1}{2}} C_y C_x^{\frac{1}{2}})^{\frac{1}{2}}). \tag{1}$$

As in [3], we simplify the above formulation to:

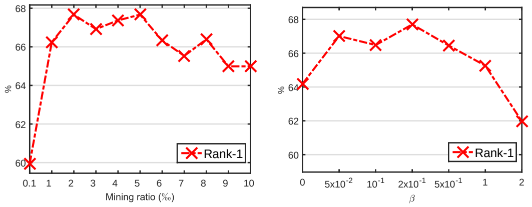
$$W_2(X,Y)^2 = \frac{1}{2}[||\mu_x - \mu_y||_2^2 + ||\sigma_x - \sigma_y||_2^2], \tag{2}$$

where σ_x and σ_y are the standard deviation vectors of X and Y, respectively.

3. Further implementation details

We resize the images to 384 × 128 with random crop and horizonal flip in the training. In the testing we do not

*Corresponding author



(a) mining ratio

(b) β

Figure S2. Evaluation on mining ratio and β on the Market-1501.

use any data augmentation. We use the ResNet-50 [2] as our backbone network which produces a 2048-D feature embedding. We remove the last ReLU nonlinear so that the feature embedding would not be cut to a half hypersphere [6]. We set the learning rate to 0.0002. We train our model by SGD with the moment of 0.9 for 20 epoches. We decrease the learning rate by a factor of 0.1 after 12 epoches, and then we further decrease the learning rate by a factor of 0.1 after 18 epoches, following the ResNet-50 training scheme [2]. We use four Titan X GPUs and the total training time is about 10 hours. We set S adaptively within each batch to ensure we always have L_{MDL} (Eq. (4) in our main manuscript) computed in each batch, and keep T globally updated [5]. To facilitate end-to-end training we compute the vectors μ_v and σ_v (Eq. (6) in our main manuscript) within each batch, and keep updating μ and σ, following the implementation of the overall mean/std tensors in the BatchNorm [5].

4. Evaluations on mining ratio and β

Analysis on the mining ratio. We show the evaluation results of the mining ratio p (Eq. (3) for the soft multilabel-guided hard negative mining in our main manuscript) in Figure S2(a). We observe that when the mining ratio is too small, the performances drop significantly because the mining procedure is way too limited to a very small portion, and thus our MAR could only mine little latent discriminative information in the unlabeled RE-ID data. The mining ratio should also not be set too large, otherwise the model mines too many uninformative easy pairs [4, 10, 7, 9, 8], also leading to the performance drop.

Evaluation on β. We show the evaluation results of β (Eq. (9) for the reference agent learning in our main manuscript) in Figure S2(b). We observe that our model learning is stable when $5 \times 10^{-2} < \beta < 5 \times 10^{-1}$, although it should not be set too large to violate the balance of the loss magnitudes in Eq. (9) in our main manuscript.

References

[1] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[3] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *TPAMI*, 2018.

[4] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[7] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[8] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[9] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[10] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[11] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.