Person Re-IDentification

Andrea Bonomi, Khouloud Ismail, Francesco Laiti, Davide Lobba, Evelyn Turri

Abstract—Person Re-IDentification (Re-ID) is the task of recognizing a person who has previously been observed over a camera network. In this report we first give an overview of the topic and present the widely used standard baseline. Then we analyze three papers: *Bag of Tricks and A Strong Baseline for Deep Person Re-identification* [1] and *NFormer: Robust Person Re-identification with Neighbor Transformer* [2], respectively released in 2019 and 2022 for Person Re-ID task, and *Multi-Domain Learning and Identity Mining for Vehicle Re-Identification* [3] related to Vehicle Re-ID task released in 2020. Finally, we explain the motivation behind our papers' choice.

1 Introduction

Person re-identification (Re-ID) aims at retrieving a specific person from a large number of images captured by a multi-camera network having non-overlapping field-of-views. This task is highly challenging due to changes in person poses, different camera views, illumination conditions, and occlusions. This topic plays an important role in building smart cities and for public safety.

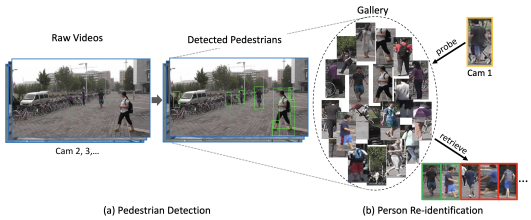

Fig. 1: An end-to-end Person Re-ID system. [4]

The procedure shown in Fig. 1 takes raw video frames and a query bounding box *Cam 1* (representing the query person-of-interest) as input. One is required to first perform pedestrian detection on the raw frames, and the resulting bounding boxes form the Re-ID gallery. Then, a standard Re-ID approach is applied. A critical aspect of this pipeline is that a better pedestrian detector tends to produce higher Re-ID accuracy, given the same set of Re-ID features [4].
This method can be extended to other object re-identification tasks, such as vehicle Re-ID.

2 Deep Person Re-ID

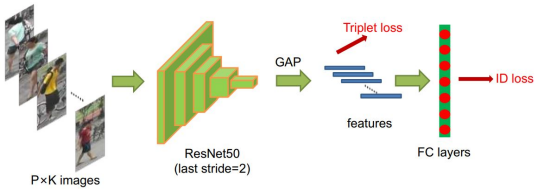

Fig. 2: The pipeline of the standard baseline. **GAP**: Global Average Pooling. **FC**: Fully Connected. **ResNet50** [5]: backbone network. **ID loss**: cross-entropy loss renamed by the authors of BoT-BS [1].

Presently, Re-ID deep learning methods dominate over hand-crafted ones, with impressive results on public datasets [4]. Deeply-learned representations provide high discriminative ability, especially when aggregated from deeply-learned part features. When working with highly varying settings across different cameras and scenarios, robust and discriminative representation learning is required.
As mentioned in BoT-BS [1], Fig. 2 provides a scheme of a widely used open-source deep learning baseline [6] adopted by the following papers .

3 Bag of Tricks

Most of Person Re-ID related works and papers design complex network structures, concatenate multi-branch features in order to encode different-level information, and therefore achieve high performances.
Despite their impressive efficiency during the training stage, these models come at a price of their low speed during the retrieval stage and their computational cost. These are major barriers to deploying them in resource-constrained settings with strict latency requirements. That urged the need to design a simple and efficient baseline for research, industry, and the community.
This baseline is designed in the "*Bag of Tricks and A Strong Baseline for Deep Person Re-IDentification*" paper [1], which performs training tricks on the standard baseline (Fig. 2) and achieves, so far, the best performance using only global features. In the next subsections, we explain each trick and why they are important.

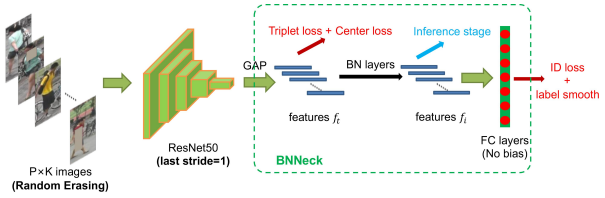

Fig. 3: Pipeline of BoT baseline [1].

3.1 Warmup Learning Rate

In Person Re-ID tasks, the learning rate plays a fundamental role. Given that the model parameters are initialized using

a random distribution, a warmup strategy is applied to bootstrap the network for better performance.
As it is shown in Fig. 4 the learning rate follows a linear increase in the first phase, and then it is reduced relying on the number of epochs.

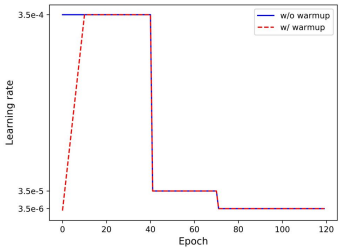

Fig. 4: Warmup Learning Rate [1].

3.2 Random Erasing Augmentation

One of the major problems in Person Re-ID is the occlusions. To address this, it has been proposed *Random Erasing Augmentation* (REA) which is a method for data augmentation. REA randomly selects a rectangle region in an image and erases its pixels with random values. It is useful to improve the generalization of the model, but in cross-domain Re-ID tasks the REA makes the performances drop.

3.3 Label Smoothing

In this paper, the cross-entropy loss is named ID loss and it is computed as:

$$L_{ID} = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, & y \neq i \\ q_i = 1, & y = i \end{cases} \tag{1}$$

where person ID is denoted as ID, truth ID label as y, ID prediction logits of class i as p_i, and N is the number of people.
Given that Person Re-ID is a one-shot learning task, it is important to prevent the Re-ID model from overfitting training IDs. To address this, it has been proposed *Label Smoothing* which is a regularization method that makes the model less confident during the training phase. In particular, it changes the q_i in the following way:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & \text{if } i = y \\ \varepsilon / N & \text{otherwise} \end{cases} \tag{2}$$

where ε is the regularization parameter.
The performance of the model can be significantly improved using label smoothing, especially if the training set is rather small.

3.4 Last Stride

The last stride is the last spatial down-sampling operation of the backbone architecture. Given the fact that higher spatial resolution always enriches the granularity of features, the last stride has been reduced from 2 to 1, in order to obtain feature maps with increased spatial size. This was possible without the addition of extra computational cost because it does not add extra training parameters.
This trick has improved consistently the performance.

3.5 BNNeck

Standard baseline, and most of the related works, combine triplet loss and ID loss to train Re-ID models. These losses constrain the same feature in the stated works, however, this paper observes the inconsistency between these two losses. ID loss constructs several hyper-planes to separate the embedding space into different sub-spaces (each subspace identifies an identity/a class). Therefore, cosine distance is the most suitable metric for the inference stage. Triplet loss enhances the intra-class compactness and inter-class separability in the Euclidean space, and Euclidean distance is the suitable metric.
The solution proposed in this paper is the design of a new structure *BNNeck* shown in Fig. 3; which adds a batch normalization (BN) layer after features (and before classifier FC layers).

3.6 Center Loss

Given that triplet loss $L_{Triplet}$ does not take into consideration the distance between each feature and their corresponding class-center, the authors of the paper proposed the *Center Loss*:

$$L_C = \frac{1}{2} \sum_{j=1}^{B} \left\| \mathbf{f}_{t_j} - \mathbf{c}_{y_j} \right\|^2 \tag{3}$$

where $\mathbf{f}_{t_j}$ is the jth feature before the BN layer, y_j is the label of the jth image in a mini-batch, $\mathbf{c}_{y_j}$ denotes the y_jth class center of deep features and B is the number of batch size.
L_C takes the entire training set into account and averages deep features of every class into a center, and penalizes distances between deep features and their corresponding class center, encouraging the model to learn widely-separated class representations.
In the end, the final loss is computed as:

$$L = L_{ID} + L_{Triplet} + \beta L_C \tag{4}$$

where L_{ID} is the ID loss in Eq. (1), $L_{Triplet}$ is the triplet loss [1], L_C is the center loss in Eq. (3), and β is the balanced weight for center loss.

4 NFormer

Most of the papers and works about Person Re-ID consider learning representations of single images, without taking into consideration the relations between multiple images of the same object. For this reason, "*NFormer: Robust Person Re-identification with Neighbor Transformer*" paper [2] is crucial, indeed the presented model learns features also from the interactions between images.
Few other works already took into consideration relations between images, but all of them were taking only a little set of images at training time, without considering the relations' results during the test time.
Including the relations between images helps either in having better results and avoiding outliers of the single image method, which usually are caused by problems such as occlusions, different dresses and different viewpoints.
To reach the described goal, the authors proposed the *Neighbour Transformer Network* (NFormer). This network aims to

model the relations among all the input images both at training and test time. With this new type of representation they also introduce a new attention module called *Landmark Agent Attention* (LAA), and a new type of softmax function called *Reciprocal Neighbor Softmax* (RNS).

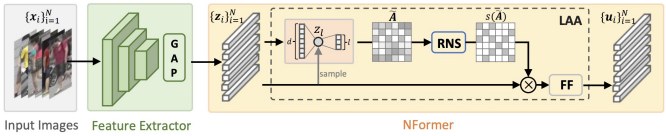

Fig. 5: NFormer Architecture [2].

4.1 Affinity matrix

The first step of the NFormer architecture is to compute an affinity matrix $\mathbf{A}$, which represents the relations between the individual representations. Matrix $\mathbf{A}$ is computed as:

$$\mathbf{A}_{ij} = \frac{K(\varphi_q(\mathbf{z}_i), \varphi_k(\mathbf{z}_j))}{\sqrt{d}} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} \tag{5}$$

where: $\varphi_q(\cdot)$, $\varphi_k(\cdot)$ are two linear projections, and they map the input representation vectors $\mathbf{z} \in \mathbb{R}^{N\times d}$ to query and key matrices $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{N\times d}$. N is the number of the input image and d is the dimension of the representation vectors. While $K(\cdot,\cdot)$ is the inner product function.
As second step, the NFormer aggregates the representation according to the affinity matrix.

4.2 LAA: Landmark Agent Attention

Computing the affinity matrix is computationally prohibitive. At this point the new attention module LAA plays a fundamental role, indeed the landmark agents allow factorizing the affinity computation into a multiplication of two lower-dimensional matrices.
Fig. 6 explains how the Landmark Agent Attention works. The query, the key and the value $(\mathbf{q}, \mathbf{k}, \mathbf{v})$ matrices are

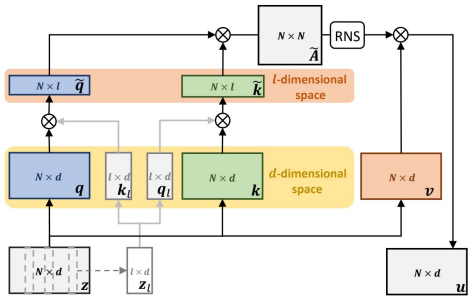

Fig. 6: LAA Pipeline [2].

obtained by three linear projections functions of the representation vector $\mathbf{z} \in \mathbb{R}^{N\times d}$. The $\mathbf{z}_l \in \mathbb{R}^{N\times l}$ are l samples from $\mathbf{z}$, which are called landmark agents, and their role is to map $\mathbf{q}$ and $\mathbf{k}$ in a lower dimensional space to obtain $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{k}}$. Multiplying $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{k}}$, the approximation of affinity matrix $\tilde{\mathbf{A}}$ is computed and it can replace Eq. (5):

$$\tilde{\mathbf{A}}_{ij} = \frac{(\mathbf{q}\mathbf{k}_l^\top)_i (\mathbf{k}\mathbf{q}_l^\top)_j^\top}{\sqrt{d}} = \frac{\tilde{\mathbf{q}}_i \tilde{\mathbf{k}}_j^\top}{\sqrt{d}} \tag{6}$$

Then the softmax is applied to $\tilde{\mathbf{A}}$, in order to obtain the attention weights. Finally, the multiplication between the weights and the matrix $\mathbf{v}$ is computed, and the final output $\mathbf{u}$ is given.

4.3 RNS: Reciprocal Neighbor Softmax

Applying a general softmax, the matrix obtained from the application of the softmax to $\tilde{\mathbf{A}}$ tends to be very noisy and dispersed. For this reason, the RNS was introduced: its role is to reduce the noisy interactions with irrelevant individuals; a number k of neighbors is set, and the remaining values are set to 0.

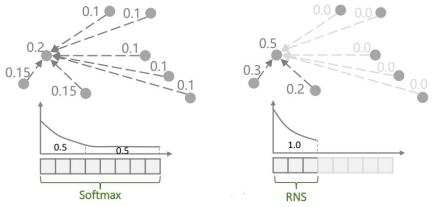

Fig. 7: Difference between a general Softmax and RNS [2].

5 Vehicles Re-ID

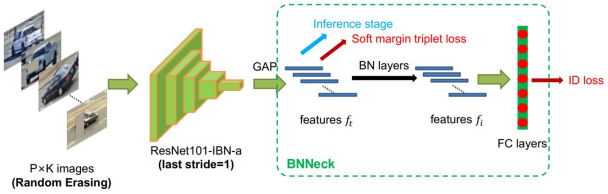

Fig. 8: Vehicle Re-ID framework [3].

S. He et al. [3] introduced a solution for the *"AI City challenge"* workshop at CVPR2020. It aims to re-identify target vehicles in images and video without knowing the license plate. In this paper they resolve two main tasks: Domain adaptation on synthetic vehicle images and Vehicle Re-ID. We investigate only the second task. In particular, they use the algorithm proposed by L. Hao et al. [1] and apply it without any change to the Vehicle Re-ID task obtaining very good performances. They studied later some changes to improve the algorithm. There are three major changes in the Vehicle Re-ID framework (Fig. 8) w.r.t. the original Bag-of-Tricks framework (Fig. 3): (1) The ResNet architecture used here is different; (2) The inference stage is moved before the Batch Normalization layers; (3) Instead of using the Triplet loss in addition to a Center loss they only use a Soft margin Triplet loss.

6 Papers' choice motivations

Starting from Person Re-ID task, we want to transfer the concept to the Vehicle Re-ID task. We studied how Vehicle Re-ID is implemented starting from an already strong baseline on Person Re-ID [1]. Now we aim to do Vehicle Re-ID with the algorithm presented by H. Wang et al. [2] by changing the dataset and bringing similar changes of Section 5 to the architecture in order to achieve better performances.

REFERENCES

[1] L. Hao, G. Youzhi, L. Xingyu, L. Shenqi, and J. Wei, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 4321–4329.

[2] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, "Nformer: Robust person re-identification with neighbor transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7297–7307.

[3] S. He, H. Luo, W. Chen, M. Zhang, Y. Zhang, F. Wang, H. Li, and W. Jiang, "Multi-domain learning and identity mining for vehicle re-identification," 2020. [Online]. Available: https://arxiv.org/abs/2004.10547

[4] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016. [Online]. Available: https://arxiv.org/abs/1610.02984

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[6] X. Tong, "Open-reid." [Online]. Available: https://github.com/Cysu/open-reid

[7] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1812.01187

[8] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016. [Online]. Available: https://arxiv.org/abs/1610.02984

[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," 2017. [Online]. Available: https://arxiv.org/abs/1711.09349

[10] L. Hao, G. Youzhi, L. Xingyu, and L. Shenqi, "Bag of tricks and a strong reid baseline." [Online]. Available: https://github.com/michuanhaohao/reid-strong-baseline

[11] W. Haochen, "Nformer." [Online]. Available: https://github.com/haochenheheda/NFormer

[12] Z. Zhedong, "Pytorch reid." [Online]. Available: https://github.com/layumi/Person_reID_baseline_pytorch