Relevance Transfer: Towards Robust Distillation in Person Re-Identification

Jing Zhou, Zuliu Yang, Junting Yang and Yong Zhao

Shenzhen Graduate School of Peking University (PKUSZ) School of Electronic and Computer Engineering, PKUSZ, Shenzhen, China

Email: geenzhou@pku.edu.cn, yangjunting100@pku.edu.cn, yongzhao@pkusz.edu.cn

Abstract. Person Re-Identification (ReID) is developing rapidly these years powered by deep learning. However most existing methods with state-of-the-art performance are highly dependent on “heavy” CNN backbones, which are hard to deploy on lightweight computing devices. In this paper, we propose a novel feature distillation method called Relevance Transfer, which aims to transfer the knowledge hidden inside the relation of different feature parts from teacher network to tiny student network. And we conduct a Multi-Granularity Horizontal Partition (MGHP) strategy to make training more efficient and robust. As results on two mainstream ReID datasets show, our method achieves competitive performance with original logits Knowledge Distillation (logits KD), and gains even more aggressive performance when works along with logits KD.

1. Introduction

Person Re-Identification (ReID)[1] is a traditional task in video surveillance, which aims to match images or videos in different cameras of the same person. It gets highly improved in performance and raises much more attention these years due to the development of deep learning. Due to amounts of challenges, such as pose variations, body parts occlusion, background noise, it is necessary to train a deep enough model to solve these highly nonlinear problems. Very deep model, which stacks a ton of convolution layers, may achieve a very high accuracy, however, is hard to be deployed on these edge computation devices. To gain better trade-off between performance and computation efficiency, Knowledge Distillation[2](KD), which is widely used for model compression and acceleration, seems to be a good solution.

In this paper, we proposed a new feature distillation method named Relevance Transfer for ReID task. Differed from most existed feature distillation approaches, we find it a strong clue to build relation between different parts of pedestrian feature for ReID task. To fit the pedestrian information, we further propose a Multi-Granularity Horizontal Partition (MGHP) strategy to match cross-scaled features as well as reduce computation cost in training. Overall architecture is depicted in Figure 1. The main contributions of this work can be summarized into 3 aspects: 1) We propose relevance matrices with two kinds of similarity to model relation between different parts, and transfer the relevance knowledge via MSE Loss. 2) We propose a Multi-Granularity Horizontal Partition strategy to encode robust relevance knowledge. 3) Our Relevance Transfer method obtains competitive results with logits KD, and can work consistently with logits KD and gain even better performance.

2. Related Works
In this section, we mainly discuss some previous works related to Person ReID task and knowledge distillation.

2.1. Person Re-Identification
Person Re-Identification aims to match pedestrians across non-overlapping camera views. Profited by development of deep learning, features extracted by end-to-end trained CNNs become mainstream instead of handcrafted features. PCB[3] learns part-level deep features by horizontally partition the network feature map into different parts, and use a RPP strategy to refine boundaries of different parts. MGN[4] adds a global branch to PCB and uses multi-granularity partitions, remarkably improve the Rank-1 score to 95.9. He et al. [5] release an open-source pytorch toolbox called FastReID, including multiple baseline models with solid tricks. The teacher network in our paper is implemented with FastReID.

2.2. Knowledge Distillation
Model pruning, model quantization and knowledge distillation are top three most popular approaches for model compression and acceleration. Knowledge distillation was first proposed in [2], it uses softened logits of teacher’s output to transfer relevance hidden in different labels, thus the student network has to mimic teacher’s output meanwhile predicts the ground truth label. To gain more information from the teacher network, some methods furtherly use intermediate features from the teacher network to improve the student network performance. FitNet[6] aims to train a deep but thin student network hinted from a shallow but wide teacher network. Guided layers from the student are penalized to directly predict the output of hint layers from the teacher. However, existing feature distillation methods are not specific for person ReID task, and have a shortcoming in common: student’s performance gets limited improvement when feature channels from student’s hidden layers don’t match that of teacher’s. To deal with these problems, we take more consideration of the pedestrian information and propose a new feature distillation approach, which there is no need to match feature channels at all.

3. Approach
Our purpose is to train a tiny enough model with less performance degradation from big teacher model. It is known that pedestrian features in different spatial parts always have some robust relation, which can be a strong clue to identify a pedestrian. In this paper, we conduct relevance matrices to encode the relation between different parts. Furtherly, we apply a Multi-Granularity Horizontal Partition(MGHP) strategy to gain a better trade-off between performance and computation/memory cost.

3.1. Relevance Matrices
Given a feature $F \in \mathbb{R}^{N \times C}$, where $N$ means there are $N$ different spatial partitions, and $C$ is the channel dimension of $F$. In order to model the relevance of different parts, we usually build a matrix $A_{N \times N}$, which each element $a_{i,j}$ of it can be formulated as:
\[ \alpha_{i,j} = \text{Sim}(f_i, f_j) \]  \hspace{1cm} (1)

Where \( f_i, f_j \in \mathbb{R}^C \) are features at \( i \)-th and \( j \)-th \((0 \leq i, j < N)\) partition of \( F \), and \( \text{Sim}(\cdot) \) represent a certain similarity function. Since two features usually share similar semantic information when the angle between them is small, we first choose cosine similarity to measure the relevance, denoted as:

\[ \text{CosSim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \] \hspace{1cm} (2)

Where \( \cdot \) computes inner product of two vectors, and \(|\vec{a}|\) represents the norm of vector \( \vec{a} \), \( \text{CosSim}(\vec{a}, \vec{b}) \in (-1, 1) \). However, it is not enough if we ignore the intensity relation between two spatial parts, and the norm of a feature can represent its intensity. Thus a norm similarity can be formulated as:

\[ \text{NormSim}(\vec{a}, \vec{b}) = \frac{2|\vec{a}||\vec{b}|}{|\vec{a}|^2 + |\vec{b}|^2} \] \hspace{1cm} (3)

\( \text{NormSim}(\vec{a}, \vec{b}) \in (0, 1) \) scores high when two features are similar in norm. Since we use two kinds of similarity to encode relevance information, actually we need two relevance matrices denoted as \( A_{\text{cos}} \) and \( A_{\text{norm}} \), which the elements in \((i,j)\) are \( \alpha_{i,j}^{\text{cos}}, \alpha_{i,j}^{\text{norm}} \) respectively:

\[ \alpha_{i,j}^{\text{cos}} = \text{CosSim}(f_i, f_j) \] \hspace{1cm} (4)

\[ \alpha_{i,j}^{\text{norm}} = \text{NormSim}(f_i, f_j) \] \hspace{1cm} (5)

Finally, to transfer the relevance knowledge, we just apply MSE Loss to encourage the student network to predict teacher’s relevance matrices.

3.2. Multi-Granularity Horizontal Partition

It is intuitive that we can just treat every pixel in a feature as a cell to partition the feature, while it also raises huge memory and computation cost with \( O(H^2W^2C) \), besides there are many background or junk pixels in a feature which contribute nothing to identify a person, just a waste of computation and memory. Referring to [3][4], the semantic information varies little within a horizontal strip, thus we can horizontally partition the feature into strips, and apply average pooling for each to coarsely represent the original features. Result feature \( F \in \mathbb{R}^{H \times C} \) greatly scales the cost to \( O(H^2C) \). In order to represent body parts with different scales, we furtherly partition the feature hierarchically after horizontally strip pool, and concatenate them together. The relevance matrices built from multi-granularity features can match cross-scaled parts, which works more robustly in knowledge transfer. The whole routine of MGHP strategy can be checked in Figure 2.

4. Experiments
4.1. Datasets

Market1501[7] maybe is the most popular image-based Re-ID dataset, which has 32,668 images in total and includes 1501 pedestrian IDs from 6 non-overlapped camera views. The whole dataset is divided into two parts for training and testing respectively. Training set contains 12,936 images of 751 pedestrian IDs, testing set contains 3,368 query images and 19,732 gallery images of 750 pedestrian IDs. DukeMTMC-reID is a subset of DukeMTMC dataset. It contained 36,411 images of 1,812 pedestrian IDs collected from 8 cameras. However, only 1404 IDs can be observed across more than 2 cameras. We also divide the whole into two parts, the training set includes 702 IDs with more than 2 camera views, and testing set owns the other 702 IDs as well as 408 IDs(as distractors) collected from only 1 camera view. We follow Rank-1 accuracy, mean Average Precision, and mean Inverse Negative Penalty [8](mINP) to evaluate the performance in both of these two datasets.

4.2. Implementation Details

The teacher network we choose is a ResNet50-IBN model implemented with FastReID[5], which Rank-1 accuracy is 95.2 and mAP score is 88.1 on Market1501. And we choose a ResNet18-IBN model as the student baseline network, which has a big gap in performance with the teacher network, only scores 91.8 in Rank-1 accuracy and 78.1 in mAP. The initial learning rate is set to 0.0003, and the learning rate decreases with a factor 0.1 every 40 epochs. Besides, a linear warm-up strategy is applied in first 10 epochs. All our experiments are trained with one NVIDIA TITAN X GPU.

4.3. Comparison to Original KD Method

The results of Market1501 and DukeMTMC-reID shown below will prove the validity of our proposed method.

| Models          | Market1501 | DukeMTMC-reID |
|-----------------|------------|---------------|
|                 | Rank-1    | mAP | mINP | Rank-1 | mAP | mINP |
| Student Baseline| 91.78 | 78.03 | 43.4 | 84.61 | 71.02 | 31.45 |
| Logits KD       | 94.18 | 85.41 | 58.57 | 87.70 | 76.62 | 40.37 |
| RT              | **94.74** | 83.89 | 54.55 | 87.07 | 74.03 | 36.22 |
| RT + Logits KD  | 94.66 | **86.01** | **60.7** | **88.46** | **77.22** | 40.18 |
| Teacher         | 95.20 | 88.13 | 65.28 | 89.18 | 79.61 | 45.21 |

From Table 1, it is easy to find that our Relevance Transfer method gains a great improvement, which outperform the baseline model 2.96%, 5.86%, 11.15% in Rank-1 accuracy, mAP, mINP respectively on Market1501 dataset, and 2.46%, 3.01%, 4.77% on DukeMTMC-reID dataset. Though scores a little bit lower than logits KD in mAP and mINP, our method still achieves competitive results in Rank-1 accuracy, even 0.56% higher than logits KD on Market1501 dataset. Most importantly, when we train the student model with both of those two KD methods, we can obtain an even stronger student model, which gets all evaluation criteria improved from the logits KD method. As the results shown, our Relevance Transfer is still a very effective way and can furtherly extract more dark information in intermediate features, so that improve the performance of logits KD.

4.4. Ablation Studies

The spotlight of our proposed method is contributed to two main components, the relevance matrices and Multi-Granularity Horizontal Partition strategy. While the relevance matrices depend on two kinds
of similarity: cosine similarity and norm similarity. We conduct ablation experience on Market1501 dataset to verify the effectiveness of each component.

**Table 2.** Ablation results (%) on Market1501 dataset. N_Sim represents results only use Norm Similarity, C_Sim only use Cosine Similarity and CS means using Cross-Scaled feature mapping. Value in bold fonts still means the best student model in each criterion.

| Models                | Market1501 |
|-----------------------|------------|
|                       | Rank-1     | mAP   | mINP  |
| Student Baseline      | 91.78      | 78.03 | 43.4  |
| N_Sim                 | 93.26      | 81.9  | 51.24 |
| C_Sim                 | 93.62      | 82.83 | 53.51 |
| N_Sim+C_Sim           | 93.85      | 83.53 | 54.32 |
| N_Sim+C_Sim+CS        | **94.74**  | **83.89** | **54.55** |
| Teacher               | 95.20      | 88.13 | 65.28 |

From Table 2, we can observe that the performance of each criterion gains improvement after a new component added. When we only use Norm Similarity or Cosine Similarity to form a relevance matrix, all criteria get a considerable improvement, especially for Cosine Similarity. If we use both two similarities, mAP and mINP elevate fairly. Finally, we add the cross-scaled feature mapping strategy (concatenate multi-granularity features to generate similarity matrix), Rank-1 score gets significant increase.

5. **Conclusions**
In this paper, we propose a novel feature distillation method with a competitive performance itself, and there is no conflict between these two KD methods, thus our method can work in a perfect harmony with logits KD. The process of Relevance Transfer is kind of similar with how human vision system works: we always learn something new by finding the relation of what we saw before or learnt before. It should be very funny to apply our proposed method to other computer vision tasks.

6. **References**
[1] Zheng L, Yang Y, Hauptmann A G. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.
[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
[3] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480-496
[4] Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 274-282.
[5] He L, Liao X, Liu W, et al. Fastreid: A pytorch toolbox for general instance re-identification[J]. arXiv preprint arXiv:2006.02631, 2020, 6(7): 8.
[6] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.
[7] Zheng, Liang, et al. "Scalable person re-identification: A benchmark." Proceedings of the IEEE international conference on computer vision. 2015.
[8] Ye M, Shen J, Lin G, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.