Relation Aware Attention for Person Re-identification

Junyu Song\textsuperscript{1,a}, Kaifang Li\textsuperscript{1,b}, Guancheng Hui\textsuperscript{1,c} and Miaohui Zhang\textsuperscript{1,d}*. \\
\textsuperscript{1} School of Artificial Intelligence, Henan University, Kaifeng, Henan, 475004, China \\
\textsuperscript{a}email: sjy@henu.edu.cn, \textsuperscript{b}email: lkf59553@163.com, \textsuperscript{c}email: hgc@henu.edu.cn \\
*Corresponding author: demail: zhmh@henu.edu.cn

Abstract. Person re-identification (Re-ID) means matching people across different camera views based on different locations. It is challenging for person images where there are background clutter, pose variations, illumination changes, etc. Attention mechanisms have become attractive for person re-identification algorithms as they aim at strengthening discriminative features, which accord with the purpose of person re-id, i.e., learning discriminative features for different pedestrians. Previous approaches mostly learn attention by using local convolutions which have limited receptive fields. In this paper, an Relation Aware Attention (RAA) module is proposed to address this issue. RAA infers attention maps along two dimensions, channel and spatial, then they are multiplied to the feature as the output map. For each feature position, RAA harvests the pairwise relationship with others as its response. Furthermore, in order to grasp the structure information of global scope and the local appearance information, we stack the relations and the feature to learn the final attention with a convolutional model. We design the experiment and compare it with the existing benchmark. The results show that our attention model can increase the ability of feature representation.

1. Introduction

Person re-identification (Re-ID) is considered to be a pedestrian retrieval problem. Given a query person image and search for the same person across non-overlapping cameras\cite{1} which has aroused widespread interests in both academia and industry. Due to the increasing number of security cameras and the demand for security monitoring. For example, searching for missing elderly and children in crowded places such as railway stations or scenic spots to help public security organs quickly identify suspects and track them. Re-ID searches the specific pedestrian in gallery through a probe image, which is usually combined with pedestrian detection and pedestrian tracking.

Re-ID is a challenging task as shown in Figure 1, pedestrian images often suffer from different poses, complex background clutter, varying resolution, occlusion, etc. Therefore, how to learn a robust distinguishing feature is the key to improve the performance of person reidentification.

For this reason, most Re-ID algorithms focus on learning pedestrian features that are distinctive and robust to meet the above challenges. The existing methods are roughly divided into two kinds, one is called feature learning, the other is called metric learning. Feature learning refers to solving problems by artificially designing a series of targeted features \cite{2} or different depth features \cite{3}. Metric learning attempt to minimize intra class distance and maximize inter class feature distance by mapping images to feature space, so as to improve accuracy \cite{4}. Many existing networks combine the two approaches to achieve a better result. However, because of the complex background and the great changes in the appearance of pedestrians, there are still confusing positive samples, and even through feature learning and metric learning, the probe will be more similar to negative samples.
Figure 1 Challenges of Re-ID. (a) pose challenge, (b) background clutter, (c) varying resolution, (d) occlusion. The sample images in the same box belong to the same person.

When comparing two images, people look repeatedly at areas of interest, which often contain more discriminating features [5]. Therefore, we designed spatial relation aware attention module (RAA-S) and channel attention module (RAA-C), which explore the global structural relationships on space and channel dimensions respectively. At present, many excellent attention algorithms have been proposed. Yang et al. adopted deep convolution model to obtain large receptive field [6]. [7] adopted large convolution kernel to obtain spatial attention in their model. However, these methods can not explore the relationship between nodes. Non-local block [8] proposes to use the relationship between feature points to obtain the global context relationship. Inspired by Non-local blocks, our RAA module also uses node relationships to obtain global information. at the same time, we simplify the operation and stack the relationship mask and the feature graph itself to train together.

In summary, the work of this article is mainly in the following two aspects.

- We propose to use node relationships to mine global context information and cooperate with local appearance information to learn attention.
- We have designed the relation aware attention (RAA) module, which includes spatial dimension and channel dimension, which can be easily inserted into the backbone network.

We conduct experiment to prove the effectiveness of the RAA module on the benchmark data set CUHK03 [9], Market1501 [10].

2. Approach

In this section, we propose a relation-aware attention (RAA) module, which can be easily plugged into the backbone network to get attention. In Section 2.1 we introduce our overall network framework and our main idea, and then we introduce our spatial attention (RAA-S) and channel attention (RAA-C) modules in Section 2.2 and 2.3, respectively.

2.1. CNN backbone and Main idea

**Backbone:** We usually use ResNet-50 [11] as the backbone feature extraction network in person re-identification tasks because of its effective ability to prevent the gradient from disappearing. As shown in Figure 2, in order to obtain richer features, we remove the last lower sampling layer, and then classify pedestrians through the fully connected layer of the Ni dimension, where Ni represents the number of identities (751 for Market1501, 767 for CUHK03). The ResNet-50 is pretrained on ImageNet [12].

**Idea:** Given a set \( \mathcal{Y} = \{x_i \in \mathbb{R}^d, i = 1, \ldots, N \} \) of N features with each of d dimension. The goal of attention is to learn a set \( a = (a_1, \ldots, a_N) \in \mathbb{R}^N \), which is used to weight N features according to their importance. According to the above research, global attention can achieve long-distance dependence, e.g., using full connection layer. However, the fully connected operation has a large amount of computation and will lose the structure information of the feature graph. Therefore, we calculate the relationship \( r \) (e.g., affinity) between the current \( (i^{th}) \) feature nodes and some long-distance \( (j^{th}) \) feature points, and then superimpose with the feature graph itself \( i.e., y_i = [x_i, r_i] \) to learn attention.
2.2. Spatial relation-aware attention (RAA-S)
As shown in Figure 2, spatial attention is added after the four residual blocks of ResNet-50 (including cov1, cov2, cov3, cov4). After feature extraction, the feature graph $X \in \mathbb{R}^{C \times H \times W}$ of height $H$, width $W$, and $C$ channels is obtained. The dimension of each spatial feature node is $C$. We calculate the relationship between each feature point $x_i$ and its row and column nodes $\{x_i, x_{i-1}, \ldots, x_j\}$.

Pairwise relational $r_{ij}$ (i.e. similarity) between node $x_i$ and node $x_j$ can be defined as a dot-product in embedding space as:

$$r_{ij} = f_S(x_i, x_j) = \theta_S(x_i)^T \phi_S(x_j),$$  

where $\theta_S$ and $\phi_S$ are two embedding functions composed of 1x1 convolution layer followed by batch normalization (BN) and ReLU activation function. The convolution kernel of 1x1 reduces the eigenvector from $C$ dimension to $C/8$ dimension, carries out preliminary information integration and simplified calculation. We put all the $r_{ij}$ together to form a similarity matrix $R \in \mathbb{R}^{N \times M}$. $N$ represents the number of xi $(W \times H)$, $M$ represents the number of xj $(W+H-1)$. Each row of the matrix $r_i = [r_{i1}, r_{i2}, \ldots, r_{iM}]$ represents the relationship between the ith node and other points. We use the normalization operation in the M dimension. Then we embed $x_i$ through $\psi_S$ and stack it with the relational feature $r_i$ to get $y_i$:

$$y_i = [pool(\psi_S(x_i)), r_i],$$

where $pool(\cdot)$ represents a global averaging operation along the channel dimension, reducing the dimension to 1. We get the attention mask $a_i$ from the following function:

$$a_i = \text{Sigmoid}(\psi_S(y_i)),$$

where $\psi_S$ transforms the channel dimension to 1, $\text{sigmoid}(\cdot)$ means to map a value between 0 and 1 as an attention mask. The intuitive process is shown in Figure 3.

2.3. Channel relation-aware attention (RAA-C)
We put the channel attention after the spatial attention. Therefore, channel attention can further extract distinguishing features on the basis of spatial attention. Similar to spatial attention, we treat each dimension of the tensor $X \in \mathbb{R}^{C \times H \times W}$ as a node of the $d=H \times W$ dimension, then calculate the relationship between each pair of nodes. The relationship between nodes $x_i$ and $x_j$ is defined as the dot product of the embedded space:

$$r_{ij} = f_C(x_i, x_j) = \theta_C(x_i)^T \phi_C(x_j),$$

where $\theta_C$ and $\phi_C$ are two embedding functions composed of 1x1 convolution layer followed by batch normalization (BN) and ReLU activation function. The above operation can reduce the spatial dimension
to 1/8. As illustrated in Figure 4, we use an affinity matrix $R_{e} \in \mathbb{R}^{C \times C}$ to represent all paired relationships. Similarly, we embed the feature itself $x_i$ and then stack it with the relational feature $r_i = [r_{i1}, r_{i2}, \ldots r_{iC}]$ to get $y_i$:

$$y_i = \text{pool}(\varphi_e(x_i), r_i),$$

(5)

where $\varphi_e$ represents an embedded operation, $\text{pool}(\cdot)$ represents a global averaging operation along the spatial dimension, reducing the dimension to 1. We get the final attention mask by Sigmoid function:

$$a_i = \text{Sigmoid}(\psi_C(y_i)),$$

(6)

where $\psi_C$ reduce the channel dimension to 1.

2.4. Channel relation-aware attention (RAA-C)

In order to better monitor network training, as shown in Figure 2, we use both Triple loss and ID loss. The formula is as follows:

$$L = L_{ID} + L_{Tri}$$

(7)

3. Results & Discussion

Training and testing were conducted on two benchmark pedestrian data sets: Market1501, CUHK03. The details of the dataset are summarized in Table 1. According to the usual practice, we use cumulative matching characteristics (CMC) at rank-$i$, and average accuracy to evaluate the experimental results. Rank-$i$ means the $i^{th}$ result hit in the query. The visualization results are shown in Figure 5.
Query

Rank 1-10

Figure 5 Sample ReID results on Market1501. Green borders denote the images of the same person as the query image, while the red border denote the images of different persons.

Table 1 Detailed statistics of Market1501, and CUHK03

| Data set | Cameras | Train set (ID/Image) | Test set (ID/image) |
|----------|---------|----------------------|---------------------|
|          |         | Gallery              | Query               |
| Market1501 | 6       | 751/12,936           | 750/19,732          | 750/3368            |
| CUHK03    | 2       | 767/7365             | 700/5332            | 700/1400            |

3.1. Implementation details

All the pedestrian images are resized to 256x128. In addition, we also use random clipping, horizontal flipping, and random erasure. We use Adam optimizer to optimize our model. In the first 20 epochs, in order to prevent network divergence, we set the learning rate to $8 \times 10^{-6}$, then increase the learning rate to $8 \times 10^{-4}$, and multiply the learning rate by 0.5 every 40 epochs. All experiments are implemented with PyTorch 1.6 on 2 NVIDIA TITAN X GPUs. The results on market1501 and CUHK03 benchmark datasets are shown in Table 2.

Table 2 The performance between our method and others on benchmark datasets.

| Dataset | Method | Rank-1 | mAP | Rank-1 | mAP |
|---------|--------|--------|-----|--------|-----|
|         | Attention-guide | | | | |
|         | DCNN [13] | 90.2 | 75.6 | 60.5 | 60.2 |
|         | Mancs [14] | 93.1 | 82.3 | 66.5 | 60.5 |
|         | MHN6(PCB) [15] | 95.1 | 85.0 | 77.2 | 72.4 |
|         | Others | | | | |
|         | PCB+RPP [16] | 93.8 | 81.6 | 63.7 | 57.5 |
|         | MGN [17] | 95.7 | 86.9 | 68.0 | 67.4 |
|         | Ours | RAA | 95.8 | 87.8 | 76.0 | 74.6 |

It can be seen from the table that the result on CUHK03 is much worse than that on market1501, mainly because the background of CUHK03 is more complex. Our RAA model achieves the best results of mAP/Rank-1 = 75.6%/78.7% on cuhk03 dataset. Although other methods have achieved high results on the market dataset, our attention mechanism can still achieve competitive results.

4. Conclusions

In this paper, we propose RAA module, which includes spatial attention module (RAA-S) and channel attention module (RAA-C). Both of them rely on pairwise affinity between source node and target node to capture remote dependencies and then learn attention with local appearance features. The experimental results show that the attention mechanism proposed in this paper can mine the distinguishing features of pedestrian images, so as to help the task of ReID.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (62002100) and the Foundation of Henan Educational Committee (19A520002).
References

[1] Chen Y C, Zhu X, Zheng W S and Lai J H 2018 Person Re-Identification by Camera Correlation Aware Feature Augmentation IEEE Trans. Pattern Anal. Mach. Intell. 40 392–408

[2] Zhang L, Xiang T and Gong S 2016 Learning a discriminative null space for person re-identification Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December 1239–48

[3] Dong H, Lu P, Zhong S, Liu C, Ji Y and Gong S 2018 Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning Neurocomputing 307 25–37

[4] Chen D, Yuan Z, Chen B and Zheng N 2016 Similarity learning with spatial constraints for person re-identification Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December 1268–77

[5] Zhang M, Xin M, Gao C, Wang X and Zhang S 2020 Attention-aware scoring learning for person re-identification Knowledge-Based Syst. 203

[6] Yang F, Yan K, Lu S, Jia H, Xie X and Gao W 2019 Attention driven person re-identification Pattern Recognit. 86 143–55

[7] Woo S, Park J, Lee J Y and Kweon I S 2018 CBAM: Convolutional block attention module Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11211 LNCS 3–19

[8] Anon Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In CVPR, pages 7794–7803, 2018.

[9] Li W, Zhao R, Xiao T and Wang X 2014 DeepReID: Deep filter pairing neural network for person re-identification Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp 152–9

[10] Zheng L, Shen L, Tian L, Wang S, Wang J and Tian Q 2015 Scalable person re-identification: A benchmark Proceedings of the IEEE International Conference on Computer Vision vol 2015 International Conference on Computer Vision, ICCV 2015 pp 1116–24

[11] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December 770–8

[12] Jia Deng, Wei Dong, Socher R, Li-Jia Li, Kai Li and Li Fei-Fei 2009 ImageNet: A large-scale hierarchical image database Proc IEEE Comput. Vis. Pattern Recognit. 248–55

[13] Li Y, Jiang X and Hwang J N 2020 Effective person re-identification by self-attention model guided feature learning Knowledge-Based Syst. 187

[14] Wang C, Zhang Q, Huang C, Liu W and Wang X 2018 Mancs: A Multi-task Attentional Network with Curriculum Sampling for Person Re-Identification vol 11208 LNCS (Computer Vision – ECCV 2018)

[15] Chen B, Deng W and Hu J 2019 Mixed high-order attention network for person re-identification Proceedings of the IEEE International Conference on Computer Vision vol 2019-October pp 371–81

[16] Sun Y, Zheng L, Yang Y, Tian Q and Wang S 2018 Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol 11208 LNCS pp 501–18

[17] Wang G, Yuan Y, Chen X, Li J and Zhou X 2018 Learning discriminative features with multiple granularities for person re-identification MM 2018 - Proceedings of the 2018 ACM Multimedia Conference pp 274–82