International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Improved Split-Attention Architecture Based on Circle Loss for Person Re-Identification

Zongjing Cao[1], Hyo Jong Lee*[2]

Submitted: 26/09/2020 Accepted : 08/02/2021

Abstract: Person re-identification aims to match pedestrian images across multiple surveillance camera views. It is still a challenging task due to the partial occlusion of pedestrian images, variations in the illumination of surveillance cameras, and similar appearances of pedestrians and so on. In order to improve the representation ability of pedestrian features extracted from the convolutional neural networks, in this paper, we proposed an improved split-attention architecture for person re-identification. Specifically, we first divide the feature map into two sub-groups and then split the features in each subgroup into three more fine-grained sub-feature maps. Moreover, in order to minimize the inter-class similarity and maximize the intra-class similarity, we use circle loss and identification loss to optimize our network together. Circle loss makes the similarity scores learn at different paces, which benefits deep feature learning. The circle loss not only makes the model have higher optimization flexibility but also makes the convergence target of the model more definite. Unlike many methods that use complex convolutional neural networks to represent pedestrian feature maps in a layer-wise manner, our proposed method improves the representation ability of pedestrian features at a more fine-grained level. We evaluated the performance of our proposed network on two large-scale person re-identification benchmark datasets Market-1501 and DukeMTMC-reID. Experimental results show that the proposed split-attention network outperforms the state-of-the-art methods on both datasets with only using pedestrian global features.

Keywords: Deep learning, Convolutional neural networks, Person re-identification, Circle loss

1. Introduction

Person re-identification (Re-ID) is usually considered as a subproblem of image retrieval. Its purpose is to retrieve interested pedestrian images from the galleries, which are captured by multiple surveillance cameras at different times and occasions [1]. Compared with manual retrieval of a query pedestrian image from a set of images taken by multiple surveillance cameras, the automatic pedestrian Re-ID system can save a lot of manual labor. Person Re-ID has become one of the hot research topics in the field of computer vision, due to its application and research significance in human-computer interaction systems and visual surveillance. However, person Re-ID is still a challenging task due to the ambiguity in the visual appearance of pedestrian images, the variation of pedestrian's posture, the clutter of the background, and the change of the illumination of the surveillance camera and so on. At present, the methods to solve the task of person Re-ID mainly focus on two key tasks: (1) learning the discriminative features of the robustness of pedestrian images. (2) learning the similarity measure of two pedestrian images features. The approach of pedestrian feature extraction can be roughly divided into two categories, namely hand-crafted features and features extracted based on convolutional neural networks (CNN) learning. Before deep learning-based methods are applied to person Re-ID tasks, traditional hand-crafted algorithms have developed many methods to extract features of pedestrian images. Color, color histogram, and texture are the most used features in the traditional hand-crafted system. In recent years, with the development of deep learning technology, person Re-ID based on deep learning methods has become more and more popular. Currently, research on person Re-ID based on deep learning methods are mainly focused on improving the representation ability of extracting pedestrian features from CNN models. However, many recent person Re-ID models are usually built using ResNet [2] or Inception [3] as the backbone network. Since these models were originally designed to solve image classification tasks, they may not be suitable for person Re-ID tasks.

In this paper, we proposed an improved split-attention architecture based on ResNeSt blocks [4] for the person Re-ID task. Specifically, in each residual module, we divide the feature map into two cardinal groups, the features in each cardinal group are split into three independent sub-feature maps. Among them, the feature representation of each cardinal group is decided by the weighted combination of its split sub-feature maps. Moreover, in order to get the optimal global constraints, we combined identification loss and circle loss [5] together to train our proposed split-attention network. The main contributions of the paper are summarized as follows:

(1) We proposed a new improved split-attention network optimized by the identification loss and circle loss for the person Re-ID task. Unlike many latest methods to concatenate multi-branch features or design complex network structure, our proposed network only uses the global features of pedestrians.

(2) We evaluate the performance of our proposed split-attention network on two large-scale benchmark person

[1] *Computer Science and Engineering, Jeonbuk National University, Jeonju – 54896, KOREA, ORCID ID: 0000-0001-9715-0619*
[2] *Computer Science and Engineering, Jeonbuk National University, Jeonju – 54896, KOREA, ORCID ID: 0000-0003-2581-5268*
* *Corresponding Author Email: hlee@jbnu.ac.kr*

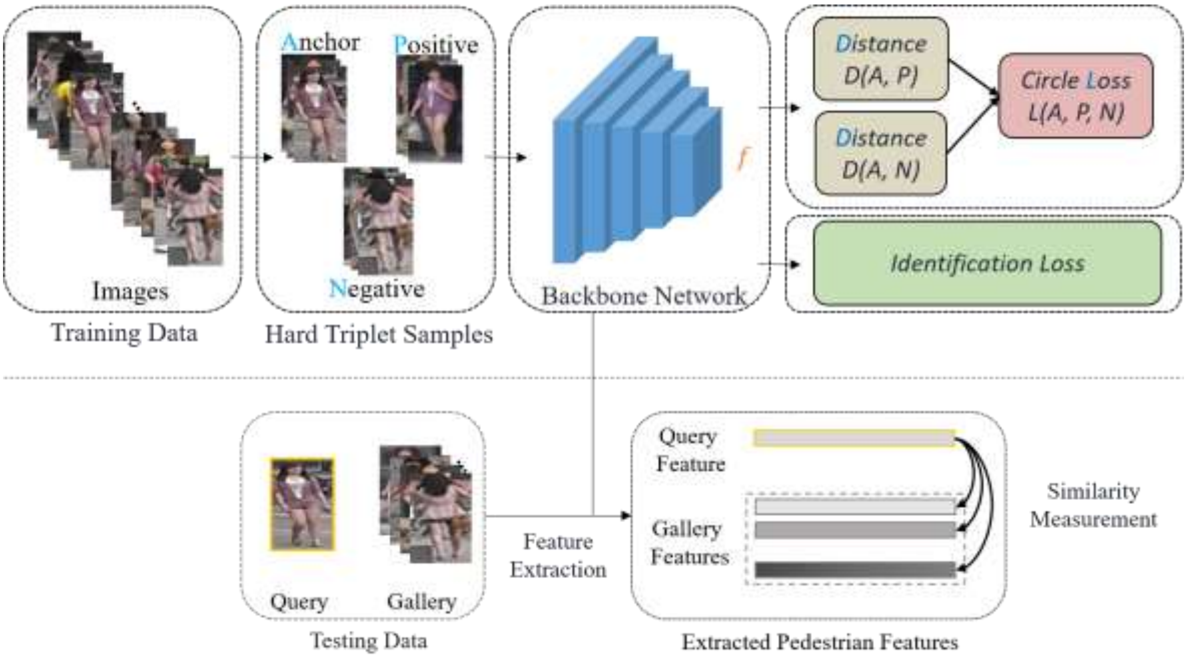

Fig. 1. The architecture of our proposed split-attention method. The upper part is the architecture of the training phase, and the lower part is the flow chart of the testing phase.

Re-ID datasets. On the Market-1501 and DukeMTMC-reID datasets, we achieve 96.10% and 89.80% rank-1 accuracy, respectively.

The rest of this paper is organized as follows: In section 2, we introduce our proposed methods and ResNeSt block in detail. In section 3, we describe our experiment and show the results. Section 4 is the conclusion part.

2. Proposed Methods

In the following subsections, we will describe the designed network, ResNeSt block, and two loss functions in detail.

2.1. Architecture of our proposed method

The goal of the person Re-ID system is to match images of the same pedestrians in the gallery captured from different surveillance cameras. The architecture of our proposed method is shown in Fig. 1. The upper part is a schematic diagram of the training process, and the lower part is the schematic diagram of the testing phase. Among them, the structure of the backbone network is the same as the ResNeSt network [4], except that the fully connected layer after the global average pooling layer is removed. The network takes triplet image as input, the images first pass through the ResNeSt network to acquire the global feature descriptors f. In the testing phase, we use this f with cosine distance metric to do the person Re-ID task.

2.2. ResNeSt Block

The ResNeSt block was proposed by Zhang et al. [4] which enables attention across feature-map groups. The internal structure of ResNeSt block is shown in Fig. 2. The split-attention block is a computing unit, which consists of a feature map group and the corresponding split attention operation.

Feature-map Group. As shown in Fig. 2, the input features of the residual block are divided into several groups. Here, we use the cardinality hyperparameter K to indicate the number of sub-feature map groups, and these sub-feature map groups are called cardinal groups. Then, the feature of each cardinal group is split into multiple smaller sub-feature groups. Here we use a radix hyperparameter R to indicate the number of splits within a cardinal group. Finally, the total number of feature groups in a ResNeSt block can be expressed as $G = K \times R$. In this work, we set $K = 2$ and $R = 3$ respectively. Then, a series of feature transformation $\{f_1, f_2, \dots f_G\}$ are adopted for the feature in each split group. The output of each split group can be expressed as $U_i = f_i(X)$, here $i \in \{1, 2, \dots G\}$. In this work, we use a 1×1 convolution layer followed by a 3×3 convolution layer as the group transformation f_i. The split attention module in each cardinal group will be described in the next section.

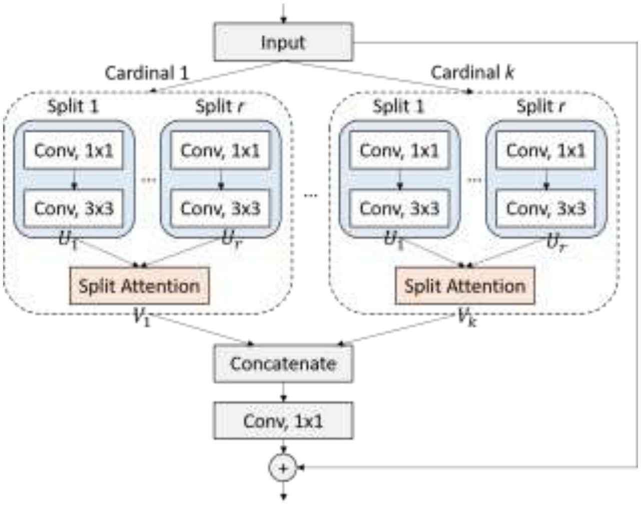

Fig. 2. Schematic diagram of the internal structure of the ResNeSt block.

Split Attention in Cardinal Groups. As shown in Fig. 3, the representation for k-th cardinal group is $\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j$, where $k \in \{1, 2, \dots K\}$. A weighted fusion of each cardinal group representation V^k is aggregated using channel-wise soft attention, where each feature-map channel is generated using the weighted combination over splits. Finally, the c-th channel of each cardinal group can be calculated as follows:

$$V_c^k = \sum_{i=1}^{R} a_i^k(c) U_{R(k-1)+i}, \tag{1}$$

where $a_i^k(c)$ indicates an assignment weight, which can be defined as:

$$a_i^k(c) = \begin{cases} \frac{\exp(G_i^c(s^k))}{\sum_{j=0}^{R} \exp(G_i^c(s^k))} & \text{if } R > 1 \\ \frac{1}{1+\exp(-G_i^c(s^k))} & \text{if } R = 1 \end{cases} \tag{2}$$

where mapping G_i^c determines the weight of each split for the $c-th$ channel based on the global context representation s^k. In this work, we use two fully connected layers with ReLU activation to parameterize the attention weight function G.

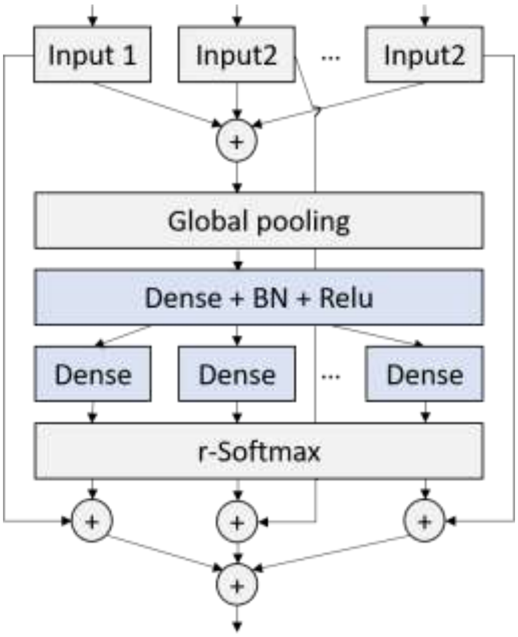

Fig. 3. Split-Attention within a cardinal group.

ResNeSt Block. As shown in Fig. 2. the cardinal group representations V^k are concatenated along the channel dimension, which can be expressed as $V = Concat(V^1, V^2 \dots V^k)$. Then sent V to another 1×1 convolution layer to fuse feature together as the output of the final ResNeSt block.

2.3. Loss function

Identification loss: The SoftMax with cross-entropy loss is usually used for identity prediction, which can by defined as:

$$L_{identification} = -\sum_{i=1}^{n} p_i \log(\hat{p_i}), \tag{3}$$

where $\hat{p}$ indicates the output of predicted probability, p indicates the target probability, and n indicates the number of samples during the training.

Circle loss: Circle loss was proposed by Sun et al. [5]. The difference from triplet loss is that circle loss simply re-weight each similarity to highlight the less-optimized similarity scores. Given a single sample x in the feature space, we use K and L to indicate within-class similarity scores and between-class similarity scores associated with x, respectively. We denote these similarity scores as s_p^i $(i = 1,2,..,K)$ and s_n^j $(j = 1,2,..,L)$, respectively. Circle loss aims to maximize the within-class similarity s_p^i and minimize the between-class similarity s_n^j. Then, the circle loss can by formulated as:

$$L_{circle} = log\left[1 + \sum_{j=1}^{L} exp(\gamma a_n^j s_n^j) \sum_{i=1}^{K} exp(-\gamma a_p^i s_p^i)\right], \tag{4}$$

where a_n^j and a_p^i are non-negative weighting factors.

Total Loss: In order to mutual benefit from identification loss and circle loss, the total loss of our model is the sum of two losses, which can be formulated as follows:

$$L = L_{identification} + L_{circle}, \tag{5}$$

3. Experimental Results

In this section, we will evaluate our proposed method on two large-scale person Re-ID benchmark datasets Market-1501 and DukeMTMC-reID and show the results of comparing our proposed model with other state-of-the-art methods.

3.1. Datasets

The Market-1501 dataset was collected in front of a supermarket on the campus of Tsinghua University [6]. The datasets contain 32,668 annotated bounding boxes of 1,501 subjects. Among them, 12936 images from 751 identities are used for the training set, 19732 images from 750 identities for the testing set, and 3368 images from other 750 identities are sued for the query image.

The DukeMTMC-reID dataset is a subset of the DukeMTMC datasets [7] for image-based Re-ID task. It contains 1404 identities that appear in more than two surveillance cameras and 408 distractor identities that appear in only one surveillance camera. The whole datasets contain 16522 training images, 2228 queries, and 17661 gallery images.

3.2. Evaluation protocol

For person Re-ID tasks, cumulate matching characteristics (CMC) and mean average precision (mAP) are two most popular evaluation metrics. In this work, we use the accuracy of CMC at Rank-1 and mAP as our evaluation metrics. For the two benchmark datasets Market-1501 and DukeMTMC-reID datasets, we employ the evaluation package provided by [6] and [8], respectively. All our experiments are evaluated based on a single-query setting.

3.3. Implementation details

Training phase: During training, the input of our network includes $B = P \times K$ images, where P and K represent the number of instances and the number of images per instance, respectively. In this work, we set $P = 6$, $K = 8$. In the training phase, all training images are augmented with a horizontal flip, random erasing operation, normalization, and resized to 384×128. In practice, we use Adam as the optimizer and train the model for a total of 90 epochs. The learning rate is initialized to 0.00035 and decay it by 0.1 at the $40-th$ and $70-th$ epochs. The backbone model is initialized by ImageNet pre-trained models.

Testing phase: In the testing phase, we first input all the gallery images into the network to obtain the pedestrian descriptors f. After the feature descriptors of the entire gallery image are obtained, these feature descriptors are stored offline. When a pedestrian image to be queried is given, the network will extract its feature descriptor online, then calculate the cosine distance between all gallery features and the query image feature and output the final ranking result. The pipeline of the testing phase is shown in Fig. 1 (lower part).

3.4. Results and analysis

Comparison with state-of-the-art methods. In this work, we evaluated our method on two large benchmark datasets described in Section 3.1. To evaluate the performance of the proposed methods, we compared five state-of-the-art approaches: partial feature-based PCB [10] methods, attribute-based PCN-PSP [1] methods, and global feature-based IDE [9], Bot [11], TriNet [12] methods. The Rank-1 and mAP on two datasets by different methods are shown in Table 1. On Market-1501 dataset, we got 95.70% rank-1 accuracy and 88.20% mAP, while on DukeMTMC-reID, our method achieved 89.80% rank-1 accuracy and 80.30% mAP. Experimental results show that our proposed split-attention architecture outperforms other state-of-the-art Re-ID methods.

Table 1. Comparison with the state-of-the-art results on market-1501 and DukeMTMC-reID datasets.

Method	**Backbone**	**Market1501**		**DukeMTMC**	
		Rank-1	**mAP**	**Rank-1**	**mAP**
IDE [9]	ResNet-50	79.51	59.87	-	-
PCN-PSP [1]	ResNet-50	92.81	78.80	85.37	71.23
PCB [10]	ResNet-50	93.80	81.60	83.30	69.20
Bot [11]	ResNet-50	94.50	85.90	86.40	76.40
TriNet [12]	ResNet-50	84.82	69.14	-	-
Ours	**ResNeSt-50**	**96.10**	**88.20**	**89.80**	**80.30**

3.5. Instance Retrieval Results

In this section, we apply the proposed model to generic image retrieval tasks and visualize some retrieval results. The samples of retrieval results are shown in Fig. 4. The leftmost image of each row is the query image, and the right (numbered 1-10) is the retrieved and sorted result. Among them, the green frame represents the image that is correctly matched, and the red frame indicates the image that is incorrectly matched. It is observed from the retrieving samples that the proposed model is also robust to partial pedestrian occlusion.

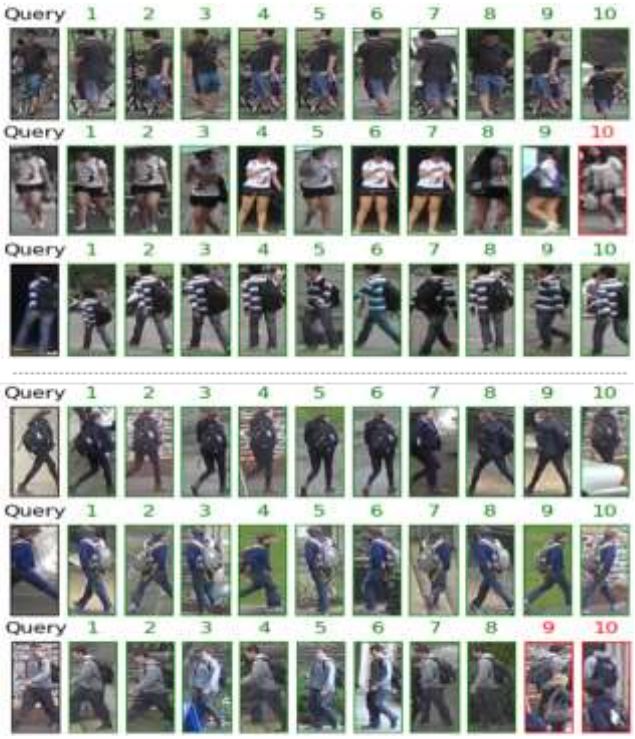
Query 1 2 3 4 5 6 7 8 9 10
Query 1 2 3 4 5 6 7 8 9 10
Query 1 2 3 4 5 6 7 8 9 10
Query 1 2 3 4 5 6 7 8 9 10
Query 1 2 3 4 5 6 7 8 9 10
Query 1 2 3 4 5 6 7 8 9 10

Fig. 4. Samples of pedestrian retrieval on Market-1501 (upper) and DukeMTMC-reID (lower) datasets.

4. Conclusion and Future Work

In this paper, we proposed an improved split-attention network based on the ResNeSt block for the person Re-ID task. To minimize the inter-class similarity and maximize the intra-class similarity, we propose use circle loss and identification loss to optimize the network together. The proposed method improves the representation ability of pedestrian features at a more fine-grained level. Unlike many latest methods to concatenate multi-branch features, our proposed network only uses the global features of pedestrians and achieves promising results. In the future, we will explore to obtain more robust pedestrian feature descriptors to further improve the performance of the person Re-ID system on a larger dataset.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (GR 2019R1D1A3A03103736).

References

[1] P. Chikontwe and H. J. Lee, "Deep multi-task network for learning person identity and attributes," *IEEE Access*, vol. 6, pp. 60801-60811, 2018, Doi: 10.1109/ACCESS.2018.2875783.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *in Proc. IEEE CVPR*, USA, 2016, pp. 770-778.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *in Proc. IEEE CVPR*, USA, 2015, pp. 1-9.

[4] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, and R. Manmatha, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020. [Online]. Available: https://arxiv.org/abs/2004.08955

[5] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," *in Proc. IEEE CVPR*, USA, 2020, pp. 6398-6407.

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," *in Proc. IEEE ICCV*, Chile, 2015, pp. 1116-1124.

[7] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *in Proc. ECCV*, Netherlands, 2016, pp. 17-35.

[8] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *in Proc. IEEE ICCV*, Italy, 2017, pp. 3754-3762.

[9] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 1, pp. 1-20, Jan. 2017, DOI:https://doi.org/10.1145/3159171.

[10] 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)," *in Proc. ECCV*, Germany, 2018, pp. 480-496.

[11] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," *in Proc. CVPRW*, USA, 2019, pp. 0-0.

[12] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv:1703.07737, 2017, [Online]. Available: https://arxiv.org/abs/1703.07737