Fusion Self-Attention Feature Clustering Mechanism Network for Person ReID

MingShou An[1][0000-0002-1180-8916], Hye-Youn Lim[2][0000 0003 0186 2430] , YunChuan He[1] and Dae-Seong Kang[2][0000-0003-0186-2430]

[1] School of Computer Science and Engineering, Xi'an Technological University, Xi'an, China
anmingshou@xatu.edu.cn
[2] Department of Electronics Engineering, Dong-A University, Busan, Korea
dskang@dau.ac.kr

Abstract. For the problem that pedestrian features can not be sufficiently extracted in person re-identification, a person re-identification model based on attention mechanism is proposed. Firstly, pedestrian features are extracted using a hybrid network combining Transformer's core multi-headed self-attention module with the convolutional neural network ResNet50-IBN-a; Secondly, an self attention mechanism is embedded to make the model of this paper more focused on the key information in the pedestrian foreground; Finally, fusing the mid-level and high-level features in the model can avoid some discriminative features loss. The experimental results show that the provide model achieves 94.8% Rank-1 and 84.5% Rank-1 on the Market1501 dataset and the DukeMTMC-reID dataset, while mAP achieves 84.9% and 65.9%.The model in this paper compares well with some of the existing person re-identification models on all the three main datasets mentioned above.

Keywords: Self-Attention, Person ReID, Feature fusion, CNN, Pedestrian foreground.

1 Introduction

Recently, because of society and the progress of artificial intelligence and the continuous increase in the flow of people in many public places, the safety hazards and crime investigation applications in public places are also becoming more and more widespread. Therefore, the popularization of intelligent monitoring equipment, which has resulted in a huge amount of surveillance video data.In order to use surveillance video data to keep people safe and improve people's quality of life, more efficient technologies is needed to process it. Person Re-identification is increasingly important in intelligent security surveillance systems, suspect tracking, intelligent people finding, etc., and is gradually becoming an important tool for maintaining public safety and social stability[1].

The rapid development of deep learning algorithms and convolutional neural networks and the improvement of computing hardware, especially the arithmetic power of graphics processing units (GPU), have enabled the use of deep learning methods for image-based person re-identification. Unlike traditional methods, deep learning-based person re-identification methods integrate the image feature extraction module and the metric learning comparison similarity module in a single model, which greatly increases the efficiency. There are still obstacles to be solved for the practical application of person re-identification, including large variations in light intensity under different cameras, inconsistent image resolution due to the distance of the captured pedestrians, and the fact that pedestrians may also be obscured by other objects such as vehicles and umbrellas.

Most current research works related to person re-identification combine global and local branches trained together to extract features from pedestrian images. However, local features often require additional models such as human pose estimation[2] or human semantic masks[3] to locate pedestrians in the images. The additional models not only increase the complexity of the model, but also the inaccurate localization will directly affect the later work. Therefore, this paper obtained effective results by only applying the global branch, including multiple attention mechanisms and feature fusion methods to extract features.

2 Method for Person ReID

2.1 Self-Attention Mechanism

The attention mechanism is used to extract representation learning to solve image misalignment problems because of its property of enhancing important features and suppressing irrelevant features. Yang et al.[4] proposed a combination of spatial attention and channel attention is proposed to learn to capture features that distinguish between the overall pedestrian image and the pedestrian part image. In addition, an interactive attention module was designed to enable the network to learn optimal weights adaptively. Li et al.[5] found that the existing methods are insufficient for soft attention. So they combined hard and soft attention mechanisms to learn important features at the region level and pixel level to solve the problem of large disparity between different graph phases of the same pedestrian, and also proposed a cross-attention interaction learning mechanism to learn global features and local features efficiently and jointly.

In this paper, the Multi-Head Self-Attention (MHSA) module in Transformer is applied to person re-identification, which can improve the performance of person reidentification by applying Multi-Head Self-Attention to replace the 3 x 3 convolution in the residual blocks of the baseline model Conv4_x and Conv5_x for person reidentification.

2.2 Feature Fusion Module

Convolutional neural network in the extraction of features, it is carried out one by one according to the residual block, and the output feature of the last residual block as a discriminator, but the final extracted features are the high-level features of pedestrians, pedestrians have the recognition of intermediate features are ignored, which has a certain impact on the recognition rate of person re-identification carried out afterwards, so the fusion of mid-level features and high-level features can make up for the shortage between.

When extracting pedestrian features, the output features of Conv3_x are 32 x 16 x 512, the output features of Conv4_x are 16 x 8 x 1024, and the output features of Conv5_x are 8 x 4 x 2048,In order to make the feature maps output by Conv3_x to Conv5_x uniform and achieve the fusion of features from multiple scales, the feature maps output by Conv3_x and Conv4_x in Figure 1 are downsampled using the maximum pooling operation.

2.3 Proposed Architecture

The general framework of person re-identification based on attention mechanism proposed in this paper is shown in Fig.1. There are mainly baseline networks, multi-head self-attention module[6], efficient channel attention module[7], and feature fusion module. In the first step pedestrian images is extracted features through the ResNet50-IBN-a[8] baseline network, where the 3 x 3 convolutional modules in the Conv4_x and Conv5_x residual blocks are replaced with multi-head attention modules, while channel attention modules are accessed after each residual block from Conv1_x to Conv5_x. The feature fusion module fuses the features of the last three residual blocks, where the output features of the Conv3_x and Conv4_x residual blocks are downsampled by maximum pooling so that they can be better fused with the output features of the Conv5_x residual block. The second step combines multiple losses such as triplet loss and cross-entropy loss for loss optimization.

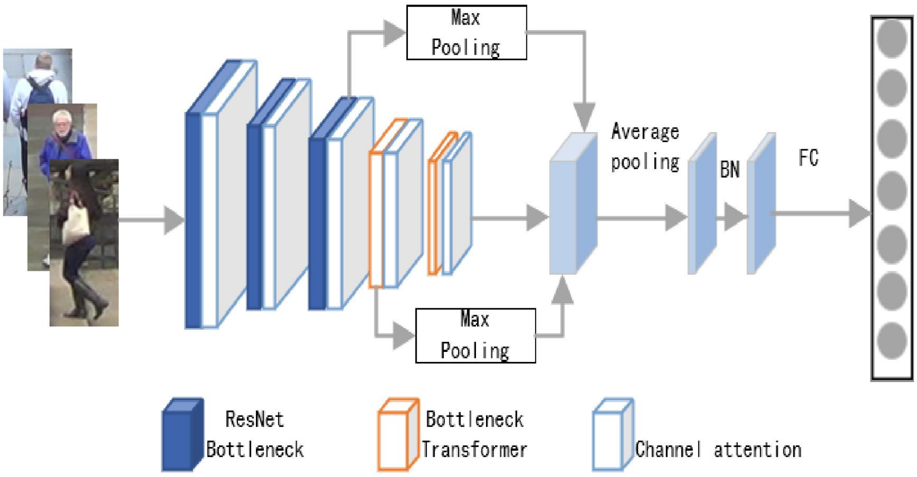

Fig. 1. General framework of Self-Attention and Feature Fusion Mechanism Network .

3 Experimental Results

To evaluate the validity of the experimental model, it was evaluated on top of three publicly available datasets, including two full-body pedestrian datasets Market-1501[9] and DukeMTMC-reID[10].

In this paper, we use the two most commonly used evaluation metrics for person re-identification, Rank-n accuracy and Mean Average Precision (mAP). Rank-n reflects the probability that the top n images with matching values among the pedestrian images to be selected are the pedestrians to be queried, and mAP integrates accuracy and recall, which can reflect the degree to which the query images are at the top of the image The mAP reflects the degree to which all correct images in the library are at the top of the retrieved list.

Table 1. Table captions should be placed above the tables.

Models	**Market1501**		**DukeMTMC-ReID**	
	mAP(%)	**Rank-l(%)**	**mAP(%)**	**Rank-l(%)**
SCPNet[11]	75.2	91.2	62.6	80.3
DCNN[12]	82.7	90.2	78.0	81.0
CGEA[13]	84.9	94.2	75.6	86.9
AlignedReID ++[14]	79.1	91.8	69.7	82.1
Our Method	82.4	94.8	65.9	84.5

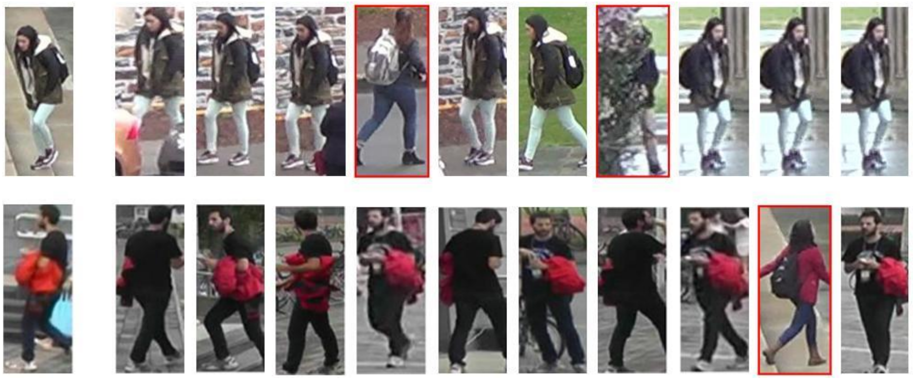

Fig. 2. The test results using our proposed method.

In Table 1, the CGEA model using graph neural network and cross-plot embedding alignment layer to jointly learn each person key point region and embed topological information achieves good results on both Market1501 dataset and DukeMTMC-reID dataset, but our model has low computational complexity and a simple structure. In the Market1501 dataset, our model improves 7.2% and 3.6%, respectively, compared to the SCPNet model Rank-1 and mAP. In the DukeMTMC-ReID dataset, the algorithm of this paper improves 3.3% and 4.2%, respectively, compared with the SCPNet model Rank-1 and mAP. The results show that the model in our paper can extract pedestrian features relatively well. Fig. 2 shows a rendering of the retrieval results on the dataset using the method proposed in this article. The first image of each row in the figure is the image that needs to be queried in the query set. The following ten pedestrian images are the sorting results obtained from the candidate set. The pedestrians marked with red borders in the figure represent those who match incorrectly. Due to explain the validity of each module added, ablation experiments were done for this purpose, and all the ablation experiments in this paper were tested with the Market1501 dataset as an example. In table 2, Baseline is the ResNet50-IBN-a network, and SAM is the Self-Attention Mechanism. Combining self attention mechanism and feature fusion for re-identification, both Rank-1 and mAP values gradually increase based on the baseline network.

Table 2. Experimental results of different block.

Models	**Market1501**	**DukeMTMC-ReID**
	mAP(%)	**Rank-l(%)**
Baseline	71.7	82.4
Baseline + SAM	76.3	89.6
Baseline + SAM + Feature Fusion	81.5	92.7

4 Conclusions

We provide this person re-identification model based on the attention mechanism, firstly, we consider that the visual Transformer has better results in the field of image processing compared with the traditional convolutional neural network, but using pure Transformer will add a large number of parameters and lead to a significant increase in computation, thus we use the core multi-head self-attention mechanism in Transformer and convolutional neural network. Secondly, a simple and effective channel attention mechanism is added to focus the model of this paper more on the important parts of the pedestrian's foreground. Finally, fusing the mid-level and highlevel features in the model avoids the loss of some distinguishing features. Experimental results on three major datasets, Market-1501 and DukeMTMC-reID, show that the performance of the proposed method is improved, and the performance metrics exceed many existing person re-identification models.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.RS-2023-00247045)

References

1. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.152−159 (2014).
2. J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang.: Attention-aware compositional network for person re-identification,"in CVPR, pp.2119–2128 (2018).
3. Z.Zhang, C Lan, W. Zeng, and Z. Chen.: Densely semantically aligned person re-identification. in CVPR, pp. 667–676 (2019).
4. Yang, F., Yan, K., Lu, S., Jia, H., Xie, X., Gao, W.: Attention driven person re-identification. Pattern Recognition, (86), pp.143-155 (2019).
5. W.Li, X.Zhu, and S.Gong.: Harmonious attention network for person re-identification. in CVPR, pp.2285–2294 (2018).
6. Srinivas, A., Lin, T. Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp.16519-16529 (2021).
7. Qilong Wang , Banggu Wu , Pengfei Zhu.: ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp.11534-11542 (2020)
8. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. Proceedings of the European Conference on Computer Vision (ECCV). pp.464-479 (2018).
9. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In International Conference on Computer Vision (ICCV), pp.1116-1124 (2015).
10. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision (ECCV), pp.17-35 (2016).
11. X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, W. Jiang, Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In Asian Conference on Computer Vision, pp.19–34 (2018).
12. Li, Y., Jiang, X., Hwang, J. N.: Effective person re-identification by self-attention model guided feature learning. Knowledge-Based Systems, 187:104832, (2020).
13. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J.: High-order information matters: Learning relation and topology for occluded person re-identification. In International Conference on Computer Vision and Pattern Recognition (CVPR), pp.6449-6458 (2020).
14. H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang.: AlignedReID++: Dynamically matching local information for person re-identification, Pattern Recognition. (94), pp.53–61 (2019).