Triplet Attention Network for Video-Based Person Re-Identification

Rui SUN†, Member, Qili LIANG†, Zi YANG†, Zhenghui ZHAO†, and Xudong ZHANG†, Nonmembers

SUMMARY Video-based person re-identification (re-ID) aims at retrieving person across non-overlapping camera and has achieved promising results owing to deep convolutional neural network. Due to the dynamic properties of the video, the problems of background clutters and occlusion are more serious than image-based person Re-ID. In this letter, we present a novel triple attention network (TriANet) that simultaneously utilizes temporal, spatial, and channel context information by employing the self-attention mechanism to get robust and discriminative feature. Specifically, the network has two parts, where the first part introduces a residual attention subnetwork, which contains channel attention module to capture cross-dimension dependencies by using rotation and transformation and spatial attention module to focus on pedestrian feature. In the second part, a time attention module is designed to judge the quality score of each pedestrian, and to reduce the weight of the incomplete pedestrian image to alleviate the occlusion problem. We evaluate our proposed architecture on three datasets, iLIDS-VID, PRID2011 and MARS. Extensive comparative experimental results show that our proposed method achieves state-of-the-art results.

key words: person re-identification, convolutional neural network, attention mechanism, deep learning

1. Introduction

Person re-identification (re-ID) is a challenging and important problem due to large alterations in viewing angles, human poses, illumination, partial occlusion, and background clutter. Recently, it plays a significant role in computer vision, and real applications includes surveillance and human behavior analysis.

Person re-ID aims to match pedestrian who appear across non-overlapping cameras. At present, existing works can be categorized into image-based methods and video-based approaches. Image-based re-ID [1]–[3] retrieval pedestrian by using multiple images. However, the information given by images is limited. In video-based re-ID, video can provide temporal information to get better pedestrian representation.

The key of video-based re-ID lies on exploiting a temporal information. In order to extract temporal feature, a lot of efforts have been made. Some works [4], [5] use temporal pooling to aggregate frame-wise features across sequences. However, they treat frames equally and ignore temporal relations of different frames. Many methods [6], [7] demonstrate that optical flow can learn motion information, which present a two-stream ConvNets to decompose the video into temporal and spatial streams. But optical flow is not suitable for occlusion problem. Recurrent Neural Network (RNN) [8] is also widely utilized to model motion cues in the entire video clip after extracting image-level features by Convolutional Neural Network (CNN). Due to numbers of fully-connected layers of it, it is complex and suffer from the difficulty in training. Attention mechanism refer to focusing on the important information and filter out the irrelevant information. Some works introduce spatial-temporal attention network [9] or attentive RNN [10], [11]. They learn the attention weight for spatial and temporal dimensions. Inspired by the attention mechanism [12], we simultaneously utilize temporal, spatial, and channel context information to enhance feature representation.

In this letter, we aim to extract discriminative and robust spatiotemporal features for video-based person Re-ID. We present a novel framework, named triple attention network (TriANet), which is shown in Fig. 1. In TriANet, we introduce residual attention subnetwork and temporal attention module. To solve the problem of background clutter, residual attention subnetwork contains channel and spatial attention module to strengthen pedestrian feature and suppress negative information. Channel attention module compute attention weight by capturing cross-dimension interaction. Spatial attention module can focus on pedestrian feature. To alleviate occlusion problem, temporal attention module...
module weight the importance of different frames and allocate the score to frames to solve occlusion problem. Extensive experiments on three challenging datasets indicate that the effectiveness and versatility of our method, which achieves state-of-the-art results.

2. Proposed Methods

2.1 Residual Attention Subnetwork

2.1.1 Channel Attention Module

Like CBAM [13], traditional ways utilize global average pooling to decompose the input tensor to one pixel per channel spatially, which leads to the loss of spatial information and lacks of correlation between spatial and channel dimension. To solve this problem, we introduce the concept of cross-dimension interaction, which use the interdependence between channel and spatial dimension to pay attention to the pedestrian features.

As shown in Fig. 2, there are two branches in the channel attention module. We capture dependencies between the (C, H) and (C, W) dimensions of the input feature respectively. Given the input feature $F \in R^{C \times H \times W}$, we let it to two branches. In the first branch, we build interactions between the height dimension and the channel dimension. Firstly, we let input $F$ rotate $90^\circ$ anti-clockwise along $H$ axis. The rotated feature is $F^\prime$, which size is $W \times H \times C$. Then, average pooling and maximum pooling operations are used in the rotated features along the spatial dimension to get two feature maps. We concatenate them together and convolve them through a $7 \times 7$ convolution layer. Finally, after the activation function, the output is rotated $90^\circ$ clockwise along $H$ axis to retain the original feature shape. In second branch, interaction is found between the height dimension and the channel dimension. Input feature $F$ is rotated $90^\circ$ anti-clockwise along $W$ axis. The rotated feature is $F^\prime_3$, which size is $H \times C \times W$. Like the first branch, we also use pooling operation, convolutional layer, and activation function to get output. Finally, we let the output rotate $90^\circ$ clockwise along $H$ axis to retain the original feature shape.

2.1.2 Spatial Attention Module

Feature map in the spatial dimension contains more effective information than the channel dimension. The spatial attention module can select spatial regions with richer pedestrian information. We use a method like the CBAM [13], and capture dependencies between the (H, W) dimensions of the input tensor to realize the generation of the spatial attention mask. This operation can generate a spatial attention mask by using the spatial relationship between the elements, and reaches the current advanced level when the number of calculation parameters is small. The structure of the spatial attention module is shown in Fig. 2.

The input of the spatial attention module is also $F \in R^{C \times H \times W}$, which is the same as the channel attention module. Firstly, we use average pooling and maximum pooling operations on the input features along the channel dimension to obtain two two-dimensional feature maps, which are connected to obtain an intermediate feature map. Then a $7 \times 7$ conv layer is used to reduce the channel of the feature map. Finally, we can get spatial attention map after sigmoid activation function. The formula is as follows:

$$A_s(F) = \sigma(f^{3 \times 7}([\text{avgpool}(F); \text{max}\text{pool}(F)]))$$

$$= \sigma(f^{3 \times 7}(F_s^{\text{avg}} + F_s^{\text{max}})),$$

Where, $\sigma$ is sigmoid activation function and $f^{3 \times 7}$ represent convolutional operation.

2.2 Temporal Attention Module

Due to issues such as lighting changes, complex backgrounds, occlusions, etc., different frames in video sequence contain different amounts of information. Pedestrians in some frames are clearer and provide more critical and effective information. These frames are called key frames, and the others are non-key frames. The key frame contains a complete and clear image of pedestrians, which can distinguish pedestrians well, while other non-key frames only provide contextual information, and their contribution to distinguishing pedestrian identities is relatively small. The function of the temporal attention module is to give more weight to key frames and reduce the weight of other non-key frames.

In temporal attention module, we design an attention generation network to extract the contextual information of the sequence, inputs are a series of image-level features maps in the network. Assume that the attention weight $a_t$ represents the importance of the $T$ frame in the sequence.

![Fig. 2](image1.png) Illustration of channel attention module and spatial attention module.

![Fig. 3](image2.png) Illustration of spatial attention module.
We use a spatial conv layer \((w, h, 2048, d_t)\) and a Fully-Connected (FC) layer \((d_t, 1)\) to calculate attention weight score \(s_t\), normalize the scores of all frames to get the final attention weight \(a_t\):

\[
a_t = \frac{\exp(s_t)}{\sum_{i=1}^{T} \exp(s_i)}
\]

Finally, weighted and sum the temporal feature vector to obtain the final video-level pedestrian feature vector \(V\):

\[
V = \frac{1}{T} \sum_{i=1}^{n} a_t f_t
\]

2.3 Visualization of Features

To verify whether our triplet attention network can adaptively localize human body without the influence of cluttered background, we visualize the receptive fields learned with our network. Grad-CAM tool is used to explain the predicted results of the model. We exploit the Grad-CAM tool to compare the baseline model and our model. In our works, Grad-CAM can identify location deemed important by the network. As shown in Fig. 4, the Grad-CAM masks of our triplet attention network are better than the baseline on covering the human body. We can clearly see that triplet attention network can adaptively localize the human body parts under complex background. Compare to baseline, our attention module more explicitly concentrates on region of the body with informative part. However, we find that our model doesn’t work fine at identifying human head. Due to the low resolution of the face, different people cannot be recognized. In summary, the visualizations further indicate the importance of employing the inter-channel and inter-spatial dependencies and temporal attention for improving feature representation.

3. Experiments

3.1 Datasets and Implementation Details

We take ablation studies to evaluate our method and compare it with existing state-of-the-arts on three prevailing video-based person Re-ID datasets, MARS [5], PRID2011 [14] and iLIDS-VID [15]. MARS contains 1,261 identities and 3,248 distractor tracks captured by six cameras. We abide by the standard of MARS and present the Rank-n accuracy and mean Average Precision (mAP). PRID2011 contains 400 sequences of 200 identities collected by two cameras. The length of each pedestrian video sequence is between 5 and 675 frames. The iLIDS-VID dataset has 600 sequences of 300 identities captured by two non-overlapping cameras. The length of each pedestrian video sequence is between 23 and 192 frames of images. Owing to changes in pedestrian perspective, posture, lighting, and serious partial occlusion, this dataset is closer to the actual situation.

All the experiments are implemented with PyTorch on Linux with GTX 1080T GPU. In terms of training, we divide each video into T blocks, and randomly sample T images as the input data. We resize the frames to 224 × 112. Common preprocessing mechanisms such as random horizontal flip, random erasing and centralization are also used as data augmentation. There are 800 epochs to train the network. We use Adam [16] with a decay rate of 0.0003 to optimize the network and decay it by 10 every 200 epochs. In addition, we set the batch size to 32. We randomly sample 8 identities, and randomly sample 4 instances for each identity. Our networks are pre-trained on ImageNet. During testing, we use the standard experimental guidelines to test the dataset. In the iLIDS-VID dataset, there are 150 pedestrians in the test set, which is half of the total. For the PRID-2011 dataset, like the method [8], this experiment uses 200 pedestrians from the 400 videos from cameras A and B to conduct the experiment. Finally, for the MARS dataset, we employ 625 pedestrian videos.

3.2 Experimental Results and Discussion

In this section, we discuss the performance of our network. All the experiments are done on the MARS dataset.

3.2.1 Effectiveness of Components

In Table 1, “no rotation” means we don’t use rotation operation in channel attention module. As we can see, using rotation can get better results. Moreover, adding attention to the spatial dimension alone can achieve better results than adding attention to the time dimension or channel dimension alone, which means person Re-ID is more dependent...
on spatial information. Superimposing the attention module of the three dimensions of channel, space and time is obviously higher than that of the network without adding attention modules or using only partial modules.

3.2.2 Influence of Sequence Length

To probe how the number of sequence length of input tracklet influences the ultimate property, we train the model with different number of frames from 1 to 8, which step size is a multiple of two. \( T = 1 \) is equivalent to an image-based model. As shown in Table 2, when \( T = 1 \), the model performance is the worst owing to limited information. When the input sequence is 2, 4, 8, the performance is relatively consistent. Their rank-1 accuracies are exceeded 84%. In addition, we can find that the performance is the best when the sequence length is 4.

3.3 Comparison with State-of-the-Art Works

In this section, we compare our approach with several recent studies on three datasets.

**MARS.** As shown in Table 3, our approach achieves 87.9% in rank-1 accuracy and 81.0% in mAP. We conduct comparative experiments with some current work. Compared with those works, our network can produce a global but highly discriminative representation to tackle background clutter, and focus on temporal features to learn more discriminative information.

**PRID and iLIDS-VID.** The comparisons on the two small datasets are shown in Table 4. We fine-tune the model, which is trained on MARS, to get the results. As we can see, our approach can achieve the comparable performance compared with state-of-the-arts.

4. Conclusion

This letter proposes a triple attention network based on video-based person re-ID. Firstly, channel attention and spatial attention are used to extract the attention map of a single frame image in the channel dimension and space dimension, and then the temporal attention module is used to obtain inter-frame dependent frames. By calculating the attention score of each frame, the video-level pedestrian feature description vector is obtained by means of weighted summation. The method in this paper focuses more on the effective pedestrian features in the video sequence and suppresses useless information, and obtains more discernible description of pedestrian features to alleviate occlusion problem, thereby improving the accuracy of re-ID. Experiments on datasets i-LIDS-VID, PRID-2011 and MARS prove the feasibility and effectiveness of the method.

Acknowledgments

This work is supported by the China Natural Science Foundation of under Grants 61471154 and 61876057. The Key Research Plan of Anhui Province -Strengthening Police with Science and Technology (202004d07020012).

References

[1] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, pp.1320–1329, 2017.

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp.1335–1344, 2016.

[3] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017.

[4] J. Gao and R. Nevatia, “Revisiting temporal modeling for video-based person reid,” British Machine Vision Conference, 2018.

[5] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “MARS: A video benchmark for large-scale person re-identification,” European Conference on Computer Vision, Amsterdam, The Netherlands, Lecture Notes in Computer Science, vol.9910, pp.868–884, Springer, Cham, 2016.
D. Chung, K. Tahboub, and E.J. Delp, “A two stream siamese convolutional neural network for person re-identification,” 2017 IEEE International Conference on Computer Vision (ICCV), pp.1992–2000, 2017.

C. Feichtenhofer, A. Pinz, and R.P. Wildes, “Spatiotemporal multiplier networks for video action recognition,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.7445–7454, 2017.

N. McLaughlin, J.M.D. Rincon, and P. Miller, “Recurrent convolutional network for video-based person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp.1325–1334, 2016.

S. Li, S. Bak, P. Carr, and X. Wang, “Diversity regularized spatiotemporal attention for video-based person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp.369–378, 2018.

Y. Liu, Z. Yuan, W. Zhou, and H. Li, “Spatial and temporal mutual promotion for video-based person re-identification,” National Conference on Artificial Intelligence, vol.33, no.1, pp.8786–8793, 2019.

Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, “See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification,” IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp.6776–6785, 2017.

D. Misra, T. Nalamada, A.U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.3138–3147, 2021.

S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, “CBAM: Convolutional block attention module,” European Conference on Computer Vision, Lecture Notes in Computer Science, vol.11211, pp.3–19, Springer, Cham, 2018.