Transfer Learning with loss combination for Person Re-Identification

Yingzhi Chen [a], Tianqi Yang

Department of Information Science and Technology, Jinan University, Guangzhou 510000, China

[a]chen_yingzhi@163.com

Abstract

The task of person re-identification is to match person images captured by different cameras, which is a important task in computer vision and is widely applied in the field of biometrics and security. Most existing re-identification method is to design a distinctive and robost representation of person, or use artificial nerual network to extract the information of person images. However, it is a extremely challenging taksk, person images have large change of pedestrian pose, camera perspective and illumination, etc. This paper is mainly based on convolutional nerual networks and transfer learing with loss combination to imporve the performence of person re-identification. The main work is summarized as follows: (1) Our work adopts ResNet structure as base model and training on ImageNet datset. (2) we combine with classification and verification loss joint supervision to learn more robust and distinct feature on person re-identificaiton dataset.

Keywords

Person Re-Identification, Loss Combination, Transfer Learning.

1. Introduction

In recent years, with the increasing emphasis on public safety issues and the rapid development of monitoring equipment, a large number of surveillance cameras are used in crowded places prone to public safety events, such as shopping malls, schools, hospitals, parks, enterprises and institutions. The appearance of surveillance cameras has greatly facilitated the detection of cases of public security agencies, such as suspects hunt, cross-scenario search, abnormal event detection, etc. Despite the increased reliability of the monitoring system, it poses a huge challenge to the management and analysis of monitoring data. At present, monitoring systems are often monitored by means of real-time camera and manual participation. Massive monitoring data is a great problem for those responsible for monitoring video surveillance. There are two reasons for this: 1) The monitoring system generates data in real time, causing a huge amount of data; 2) Real-time monitoring of data records is a scene of random changes, and the monitoring personnel are difficult to grow in the long-term observation process. The emergence of person re-identification is to overcome the shortcomings in the monitoring mechanism of human participation.

The task of person re-identification [1,2] is to study how to accurately identify people who have appeared in a particular situation in massive monitoring data, where the monitoring data is dominated by image data. The challenge of the mission is that person have complex changes in attitudes and angles in the image. In addition, the difference in illumination during shooting also makes the appearance of pedestrians change greatly. The above changes will seriously affect the performance of pedestrian recognition. Since 2012, the deep learning model represented by Convolutional Neural Network (CNN) has achieved great success in the field of computer vision. At the same time, CNN has also led research in the field of person re-identification. Compared with the traditional hand-designed person re-identification method, CNN-based pedestrian re-identification method can more effectively overcome the complex changes of pedestrians and achieve higher performance. However, pedestrians recognize that unlike other computer vision tasks (such as image classification), it is very difficult to label pedestrians, resulting in a small amount of pedestrians in existing data sets.

Thus using other datasets as auxiliary data is necessary. The transfer learning use other data as prior knowledge, but different data have gap of information. Due to this problem, we proposed a novel method to modify pre-trained CNN model and fine-tuning on the dataset of person re-identificaiton.

2. Framework for Transfer Learning

2.1 Training Process

Base Model and pre-training

The pre-training model uses the ResNet structure to extract features from the input image. The network structure has achieved good results in many filed of computer vision. This section will provide a brief introduction to the network structure. He et al. [3]proposed a deep residual network to solve the loss of information. Traditional convolutional networks or fully connected networks have more or less information loss and loss during information transmission, and also cause gradients to disappear or gradient explosions, resulting in deep network training. ResNet solves this problem to a certain extent. By directly bypassing the input information to the output and protecting the integrity of the information, the entire network only needs to learn the part of the input and output differences, simplifying the learning objectives and difficulty.We use the network training on the ImageNet dataset which is the largest dataset in computer vision to get a base model.

Fine-tuning

When we fine-tuning ResNet on person Re-ID dataset, we freeze all other layers and using a newly added softmax layer replace the original softmax layer. The reasons we modify the softmax layer are:

(1) Higher layers can be more likely to be task-specific (e.g., sensitive to general object categories rather than person identity in our case), and may have worse transferability than lower ones. In contrast, lower convolutional layers correspond to low-level visual features such as color, edge and elementary texture patterns, and naturally have better generality across different tasks.

(2) Features from higher layers are possibly contaminated by dramatic variations in human pose or background clutter due to their large receptive fields and thus not sufficiently localized for person re-id.

And we use loss combination as the new loss function of the network. Figure 1 illustrates the overall architecture of our method.

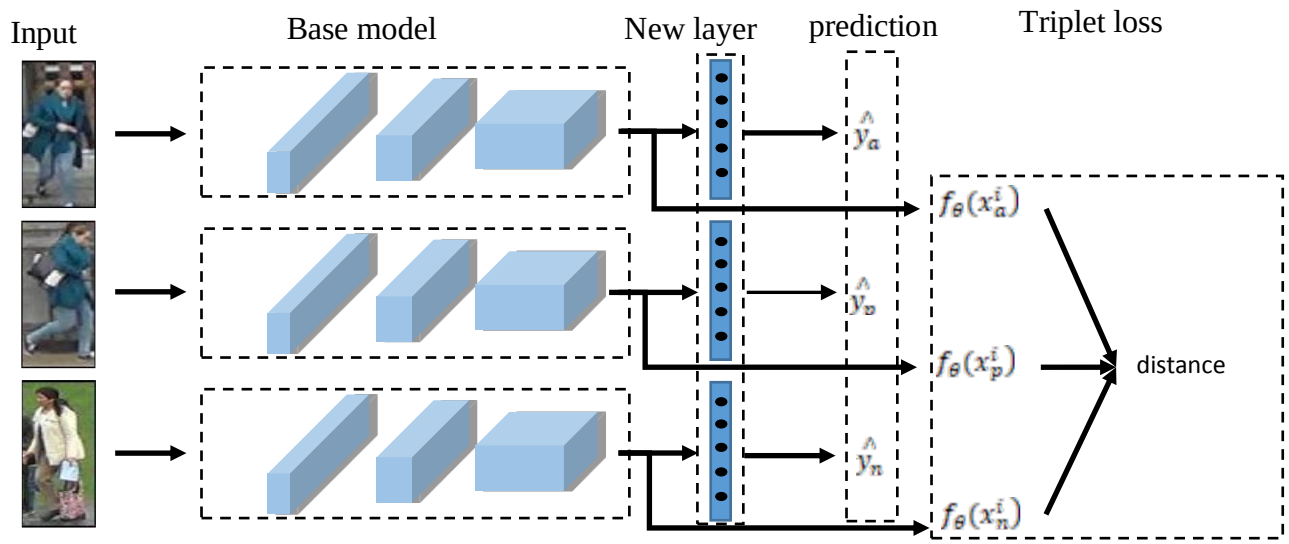

Figure 1. The overal architecture

2.2 Loss Combination

In machine learning, given a sample, the trained model is wrong for its prediction, and it should be punished. Therefore, it is necessary to define a loss function to measure the degree of inconsistency between the predicted value of the model and the true value. The loss function, also known as the

cost function, is a non-negative real-valued function. The smaller the loss function, the better the performance of the model. Thus, the goal of machine learning is to learn the number of parameters that minimize the loss function.

The last layer of the deep neural network is usually the loss layer. Therefore, an important aspect of network design is the choice of loss function. And the loss function of the neural network is more or less the same as the loss function of other machine learning models. We introduce two different loss function that is used in our neural networks.

Cross Entropy Loss

For the cross entropy loss, it aims to encourage the learned features that have similar outputs for images of the same individuals and dissimilar outputs for different individuals, which enlarge the inter-personal variations.

Which is defined as:

$$L_{CE}(\theta; x) = \sum_{i=1}^{N}[-y\log(f_\theta(x^i)) - (1-y)\log(1-f_\theta(x^i))] \tag{2}$$

Which N is the number of samples in a batch, it equals to PK .

Hard Triplet Loss

The hard triplet loss makes the distance between the matched pairs closer than that of the mismatched pairs in the learned feature space, which can effectively reduce the intra-personal variations. We project the triplet images from the original raw image space into the learned feature space through our proposed convolutional network which share the same parameter. The loss is aiming to pull the images of the same individuals closer, and meanwhile push the images belonging to different individuals far apart from each other in the projected feature space.

Suppose we have P classes and K features of each class, thus in a batch, we have PK samples.

The hard triplet loss can denoted as follows,

$$L_{trip}(\theta; x) = \sum_{i=1}^{P}\sum_{a=1}^{K}[m + \max_{p=1...K} D(f_\theta(x_a^i), f_\theta(x_p^i)) - \min_{\substack{j=1...P \\ n=1....K \\ j\neq i}} D(f_\theta(x_a^i), f_\theta(x_n^j))] \tag{6}$$

Where $D(a,b)$ denotes the distance function between a and b .

As aforementioned, in our approach, we train a convolutional network with joint contrastive and identification loss function. The joint learning objective can be described as follow

$$L(\theta; x) = \frac{1}{N}\sum_{i=1}^{N}[L_{trip}(\theta; x) + \lambda L_{CE}(\theta; x)] \tag{8}$$

The first term $L_{trip}(\theta; x)$ in (1) is the triplet loss, and N is the total number of constructed triplet training examples. The second term $L_{CE}(\theta; x)$ is the identification cost, where we have used the cross entropy as the identification cost. λ is a parameter to balance the contrastive loss and the identification loss function.

2.3 Person Matching

For person matching, we adopt the widely used Cross-view Quadratic Discriminant Analysis (XQDA). XQDA is proposed by liao et al. [7], it is formulated as a Generalized Rayleigh Quotient, and a closed-form solution can be obtained by the generalized eigenvalue decomposition.

Our method to extract CNN feature is simple, only training once one a dataset, and can straightly extract feature on any other dataset with a better performance with hand-crafted feature, the detail shown in next section. This means when our framework using on real world, we don't need collect person data and fine-tuning the network, just like hand-crafted feature, we can using the network as a feature extractor, and gain higher performance than basline.

3. Experiments

The proposed method is evaluated on the most widely used datasets in the past year for image-based Re-ID namely, VIPeR [4], CUHK-01 [5] and PRID450S [6]. The single-shot setting followed by the three datasets, where only a single query and gallery image is selected for each person. And we set multi-shot for CUHK-01, where two query and gallery image is selected for each person. The results are evaluated in form of top-k ranking accuracy.We compare the performance of our proposed method on the standard evaluation protocols for the three datasets.

We setting half of each dataset as testing data and others as training data, and ther settings same as [7].

The results are shown in Tabel 1.

Table 1 Results(rank-1) on three datasets

method	VIPeR	CUHK-01	PRID450S
Baseline	15.7	34.3	35.0
Our method	37.3	38.7	33.8

4. Conclusion

We proposed a novel method which combine with two different loss functions. It can reduce the variations of intra-personal, and increase the variations of inter-personal. Experiments on three important person re-identification datasets have convincingly demonstrated the effectiveness of our method.

And we will extend this work on two aspects. First, other models such as GoogLeNet [8] and VGGNet [9] will be used as base model to learn effective person representation. Second, other loss functions such as center loss and contrastive loss will be used.

Acknowledgements

This work was supported by the Science and Technology Project of Guangdong China under Grant No. 2017A010101036.

References

[1] Huo Z, Chen Y, Hua C. Person re-identification based on multi-directional saliency metric learning[C]//International Conference on Computer Vision Systems. Springer, Cham, 2015: 45-55.

[2] Q M B, Hu L F, Jiang J G, et al. Person re-identification based on multi- features fusion and independent metric learning[J]. Journal of Image and Graphics, 2016, 21(11): 1464-1472.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[4] Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2008: 262-275.

[5] Li W, Zhao R, Wang X. Human reidentification with transferred metric learning[C]//Asian Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 31-44.

[6] Roth P M, Hirzer M, Köstinger M, et al. Mahalanobis distance learning for person re-identification[M]//Person re-identification. Springer, London, 2014: 247-267.

[7] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2197-2206.

[8] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.