신원 확인을 위한 멀티 태스크 네트워크

조종경 [1], 이효종 [1,*]
[1] 전북대학교 컴퓨터공학부
E-mail: caozongjing@gmail.com, hlee@chonbuk.ac.kr

Multi-Task Network for Person Reidentification

Zongjing Cao[1], Hyo Jong Lee[1,*]
[1] Division of Computer Science and Engineering, Chonbuk National University
E-mail: caozongjing@gmail.com, hlee@chonbuk.ac.kr

ABSTRACT

Because of the difference in network structure and loss function, Verification and identification models have their respective advantages and limitations for person reidentification (re-ID). In this work, we propose a multi-task network simultaneously computes the identification loss and verification loss for person reidentification. Given a pair of images as network input, the multi-task network simultaneously outputs the identities of the two images and whether the images belong to the same identity. In experiments, we analyze the major factors affect the accuracy of person reidentification. To address the occlusion problem and improve the generalization ability of re-ID models, we use the Random Erasing Augmentation (REA) method to preprocess the images. The method can be easily applied to different pre-trained networks, such as ResNet and VGG. The experimental results on the Market1501 datasets show significant and consistent improvements over the state-of-the-art methods.
Keywords: Person reidentification, Verification loss, Identification loss

1. Introduction

Person reidentification (re-ID) is usually viewed as an image retrieval problem, which aims to match pedestrians from multiple cameras [1-3]. Given a person-of-interest (query), person re-ID determines whether the person has been observed by another camera [4].

Recently, the convolutional neural network (CNN) has shown potential for learning state-of-the-art feature embeddings or deep metrics [4-7]. Verification models and identification models are two major types of CNN models structures in person re-ID. The two models are different in the input, feature extraction, and loss function for training. In this work, our motivation is to combine the strengths of the verification models and identification models and learn a more discriminative pedestrian embedding. Table 1 shows that our proposed multi-task models combine the strength of the two models.

Table 1. Comparison of the Advantages and limitations of Verificati on and Identification models.

Models	Strong Label	Similarity Estimation	Re-ID Performance
Verification Models	X	O	Fair
Identification Models	O	X	Good
Our Model (Verification + Identification Models)	O	O	Good

Verification models take a pair of images (x_1, x_2) as input and predict $f(x_1, x_2) \rightarrow s$, s is a binary label, used to show whether these two inputs belong to the same person. If two inputs belong to the same person, $s = 1$, otherwise $s = 0$. Many previous works treat person re-ID as a binary class classification task [1, 8, 9] or a similarity regression task [10]. However, the major problem in the verification models is that the models use only weak re-ID labels and the annotated information is not considered.

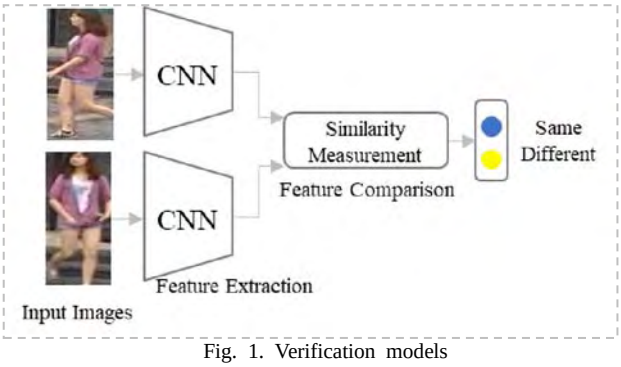

Fig. 1. Verification models

Identification models take a single image (or a batch of images) as input x and predict $f(x) \rightarrow t$, t is the predefined identity label. The cross-entropy loss is usually used following the final layer and the identification models directly learn the nonlinear functions from the input image[2]. However, the major drawback of the identification model is that the testing procedure is different from the training objective.[6] Therefore, the model does not consider the similarity measurement between image pairs, which is

problematic during the pedestrian retrieval process.

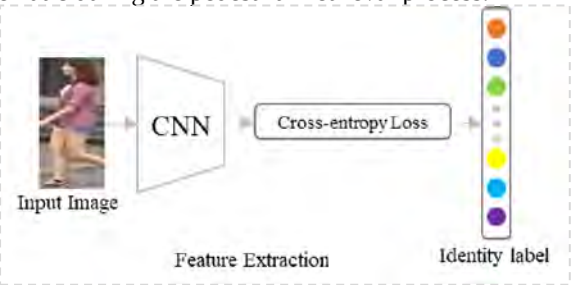

Fig. 2. Identification models

The proposed model is a Multi-Task Network that predicts person Identity labels and similarity scores at the same time. The network takes full advantage of the annotated data of the datasets and image identities [11].

2. Proposed Method

2.1 Multi-task CNN

Our network is a multi-task convolutional network that combines identification loss and verification loss. Figure 3 briefly illustrates the architecture of our proposed network. Given two images of resized to 227×227 as inputs, the multi-task network simultaneously predicts the identity label of the input images and the similarity scores.

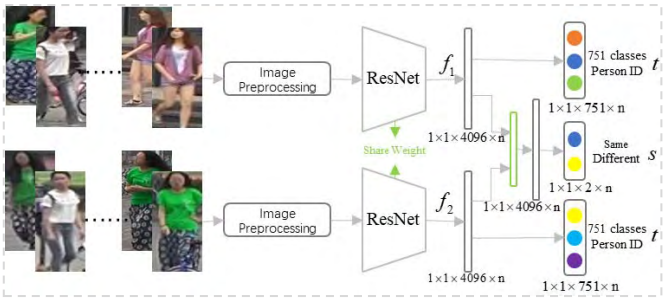

Fig. 3. Multi-Task Network models

The network consists of two ImageNet pre-trained ResNet[12] models, one square layer, three convolutional layers, and three losses.

2.2 Identification Loss

The network has two ResNet [12] models, they share weights and simultaneously predict two identity labels of the input images. Because of the number of the training identities in Market1501 datasets is 751, so the convolutional layer has 751 kernels of size $1 \times 1 \times 4096$ connected to the output f of ResNet [12] and uses softmax unit to normalize the network output. The size of the output is $1 \times 1 \times 751$. We choose the cross-entropy loss for identity prediction, which is

$$\hat{p} = soft\max(\theta_I \circ f) \tag{1}$$

$$Identify(f, t, \theta_I) = \sum_{i=1}^{K} -p_i \log(\hat{p}_i) \tag{2}$$

Here, $\circ$ is a convolutional operator, f is a $1 \times 1 \times 4096$ tensor, t is the target class, and θ_I denotes the parameters of the convolutional layer. The $\hat{p}$ is the predicted probability and p_i is the target probability, where $p_i = 0$ for all i except $p_t = 1$.

2.3 Verification Loss

As shown in Figure 3, we use a square layer to compare the features. The square layer takes f_1, f_2 as input and f_s is the output of the square layer. The square layer is denoted as $f_s = (f_1 - f_2)^2$, We treat pedestrian verification as a binary classification problem and use cross-entropy loss for predicted probability [11], which is

$$\hat{q} = soft\max(\theta_s \circ f_s) \tag{3}$$

$$Verify(f_1, f_2, s, \theta_s) = \sum_{i=1}^{2} -q_i \log(\hat{q}_i) \tag{4}$$

where the size of f_1 and f_2 is $1 \times 1 \times 4096$, s is the target class (same/different), θ_s are parameters of the convolutional layer, and $\hat{q}$ is the predicted probability. If the inputs image depicts the same person, $q_1 = 1, q_2 = 0$; otherwise, $q_1 = 0, q_2 = 1$.

3. Experiments

3.1 Dataset

Market1501: The Market-1501 dataset contains 32,668 annotated bounding boxes of 1,501 identities observed under 6 camera viewpoints [10]. The training set contains 12,936 cropped images of 751 identities, and the testing set contains 19,732 cropped images of 750 identities and distractors [10, 11]. The dataset is directly detected by the deformable part model (DPM) instead of using hand-drawn bounding boxes, which is very closer to the realistic setting. For each query, our goal is to retrieve the ground-truth images from the 19,732 candidate images.

3.2 Implementation details

Input preparation. In person Re-ID, persons in the images are sometimes occluded by other objects. To improve the generalization ability and address the occlusion problem of our models, we use the Random Erasing Augmentation (REA) method to preprocessing the images before inputting the network [11]. The images of input were been randomly cropped to 256×128 then input the network for training.

Training. The number of training epochs is 60 for our network, the initial learning rate is 0.05 and then decay learning rate by a factor of 0.1 every 5 epochs from 40 epochs. For optimizer, we use stochastic gradient descent (SGD) to update the parameters of our network [13]. We also used the same parameters to compare the results that did not use the REA to be preprocessing the images.

Testing. Given an image with 256×128 pixels, we feed

forward the image to one ResNet [12] model in our multi-task network and get a pedestrian descriptor f of which size is 4096×1 . Once the descriptors for the gallery are sets and then they have stored offline. Given a query image, the network will extract its descriptor online. We sort the cosine distance between the query image and all gallery features to obtain the final ranking result.

3.3 Performance evaluation

As shown in Table4, we evaluate our method in the single-shot setting, the proposed multi-Task network yields 92.66% rank-1 and 77.71% mAP [14] and outperforms the state-of-the-art performance.

Table 2. Comparison with the State-of-the-Art Results

Method	Rank-1	Rank-5	Rank-10	mAP
CaffeNet Baseine	35.8	65.3	77.96	42.6
VGG16 Baseline	49.1	78.4	87.2	55.7
ResNet-50 Baseline	71.5	91.5	95.9	75.8
Ours (ResNet-50)	92.66	97.50	98.27	77.71

Instance Retrieval: we apply the multi-task network to the generic pedestrian retrieval task. The results are shown in Figure 4. The image in the leftmost is the query images. The retrieved images are sorted according to the similarity score from left to right. The label of the correctly matched image is recorded in green, and the label of the false matching image is in red with a frame.

Fig. 4. Samples of pedestrian retrieval on the Market-1501 dataset

4. Conclusion

In this paper, we propose a Multi-Task Network that simultaneously considers verification and loss identification loss. It outperforms the state-of-the-art on the popular person re-ID benchmarks and shows the potential for applying it to a generic instance retrieval task. In the future, we will try to improve our network with better loss functions and data augmentation methods, and to extend the multi-task network to other applications.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (GR 2016R1D1A3B03931911).

References

[1] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, “Zero-shot person re-identification via cross-view consistency,” *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260-272, 2016.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, “Person re-identification: Past, present and future,” *arXiv preprint arXiv:1610.02984*, 2016.

[3] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification." pp. 3586-3593.

[4] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification." pp. 152-159.

[5] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, “Semantic pooling for complex event analysis in untrimmed videos,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1617-1632, 2017.

[6] 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." pp. 868-884.

[7] L. Wu, C. Shen, and A. van den Hengel, “Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification,” *Pattern Recognition*, vol. 65, pp. 238-250, 2017.

[8] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification." pp. 3908-3916.

[9] P. Natarajan, P. K. Atrey, and M. Kankanhalli, “Multi-camera coordination and control in surveillance systems: A survey,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 4, pp. 57, 2015.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark." pp. 1116-1124.

[11] Z. D. Zheng, L. Zheng, and Y. Yang, “A Discriminatively Learned CNN Embedding for Person Reidentification,” *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 14, no. 1, Jan, 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.

[13] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, “Bags of Tricks and A Strong Baseline for Deep Person Re-identification,” *arXiv preprint arXiv:1903.07071*, 2019.

[14] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification." pp. 791-808.