Compression of Person Re-identification Model Based on Depthwise Separable Convolutional Network

Liming Zhu1*, Junsheng Xiao2

1 School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
2 Henan University of Traditional Chinese Medicine, Zhengzhou, 450046, China

*Corresponding author’s e-mail: lmzhu@shu.edu.cn

Abstract. In practical application, the network depth and the number of parameters of person re-identification (re-ID) model bring great challenges to its deployment on devices with weak computing capabilities such as cloud desktop, mobile terminal and embedded terminal. To solve the above problem, this paper proposed a model compression method based on a depthwise separable convolutional network. Knowledge distillation is used to the core idea of model compression. Knowledge distillation is to transfer knowledge from a complex model to a simple model. The complex model is called the teacher model. The simple model is called the student model. This paper proposed ResNet18 based on depthwise separable convolution (ResNet18-DSC). ResNet50 is used as a teacher network and ResNet18-DSC is used as a student network. To narrow the performance gap between student and teacher networks. KL divergence loss function is used to approximate the soft label distribution of student network to that of teacher network. With a slight decrease in recognition rate, this method reduces the number of parameters by about 20 times and improves the calculation speed by about 20 times.

1. Introduction

Person re-ID is a crucial technique in urban monitoring. It has become a hot research field of computer vision. It is usually considered as a person image retrieval problem, which finds the same person image from the perspective of different cameras[1]-[2]. With the increase of network depth and scale, the recognition rate of person re-identification model is getting higher and higher, even exceeding that of human eye recognition. Complex models have better performance, but high storage space and computing resource consumption are important reasons that make it difficult to effectively apply in various hardware platforms. Therefore, the growing depth and size of convolutional neural networks have brought great challenges to the deployment of deep learning in mobile terminals. The compression and acceleration of deep learning models have become one of the key research fields in academia and industry.

Weight pruning[3] and matrix SVD[4] decomposition are the methods of early model compression, but the compression rate is far from satisfactory. Knowledge distillation is an effective method for model compression, which transfers knowledge in complex models to simple models. Deeper and more complex models are generally more expressive, but blindly increasing network depth and model parameters does not necessarily make the model more effective. The purpose of knowledge distillation is to use a model with a simpler structure and faster calculation speed to approximate the prediction effect of a complex network model as much as possible. In knowledge distillation, the complex model
is called the teacher model. The simple model is called the student model. Knowledge distillation supervises and induces student network training through soft targets of teacher network.

The main contributions of this paper are as follows:

- This paper proposed a model compression method based on depthwise separable convolutional network. ResNet50[9] is used as a teacher network and ResNet18-DSC is used as a student network. The performance of the model is improved by this method.
- With a slight decrease in recognition rate, this method reduces the number of parameters by about 20 times and improves the calculation speed by about 20 times.
- In this paper, three operating environments are set up. By comparing the operating results of the compression model in server environment, personal computer environment and cloud desktop environment, it provides a feasible reference for the operation of pedestrian re-identification tasks on devices with poor computing ability.

## 2. Related works

Weight pruning and matrix SVD decomposition are the methods of early model compression, but the compression rate is far from satisfactory. Knowledge distillation is an effective method for model compression, which transfers knowledge from a complex model to a simple model. The teacher-student model has been widely used in semi-supervised learning, model compression and knowledge distillation. It can simplify the large-scale network model without losing too much precision, which makes it possible to distribute the network to the client. The core idea of the teacher-student model is to supervise the student network through the output of the teacher network so that the student network can learn the ability to recognize the key features from the teacher network.

Knowledge distillation supervises and induces student network training through soft targets of teacher network. This is because the rough use of one-hot encoding will lose extra information about the similarity between classes and within classes. The teacher model outputs a continuous label prediction distribution for each sample, and the available monitoring information in this continuous prediction distribution is more than the traditional one-hot encoding. In addition, the method of adding loss function in the middle layer[5] can transfer the knowledge expressed in the middle layer by learning the middle layer feature map of the teacher network. The mean teacher model[6] averages model weight parameters by multiple iterative training to strengthen the supervision of unlabeled samples. Deep mutual learning[7] uses multiple student models instead of teacher models to supervise and train each other so that they can learn from each other and make progress together. Mutual mean teaching[8] design a symmetric network with hard labels and soft labels, and the two groups of networks average the model weight by learning from each other.

## 3. Proposed approach

### 3.1. Depthwise separable convolution

Different from the traditional convolution operation, ResNet18-DSC proposed in this paper adopts deep separable convolution operation. Figure 1 is the architecture of ResNet18-DSC. The conventional convolution operation is a matrix element multiplication summation of multiple convolution kernels in the receptive field and all input channels in the feature map, and the offset is superimposed. An input channel is convoluted by all convolution kernels. The number of output channels of conventional convolution operation is equal to the number of convolution kernels. In depthwise separable convolution operation is divided into two steps: depthwise convolution and pointwise convolution. In channel-by-channel convolution, a convolution kernel is responsible for a channel, and a channel is only a convolution kernel. The number of convolution kernels is the same as the number of channels in the previous feature map. The convolution kernel size in point-by-point convolution operation is $1 \times 1 \times k$, $k$ is the number of channels on the upper layer. Point-by-point convolution combines the feature map of the previous layer in the depth direction to generate a new feature map. The number of channel output equals the number of point-by-point convolution kernels. Deep separable convolution can not
only extract the dependency and correlation between regions and channels in feature map like conventional convolution but also has more efficient computational efficiency. The depth separable convolution enables the spatial and inter-channel features of the image to be calculated independently. Figure 1 is a deep separable convolution operation. The ratio of the calculation amount of the deep separable convolution operation to that of the conventional convolution operation is:

$$\frac{\text{Cal}_{DSC}}{\text{Cal}_{conv}} = \frac{1}{M} + \frac{1}{D_k^2}$$

Figure 1. The architecture of ResNet18-DSC.

3.2. The architecture of the model

In this paper, knowledge distillation is used as a model compression method. ResNet50[9] is used as a teacher network, which is parameterized as $\theta^t$. ResNet18-DSC is used as a student network, which is parameterized as $\theta^s$. Figure 2 is the overview of the proposed model compression method. In order to narrow the performance gap between the teacher network and the student network, the similarity distribution of the two predicted soft labels is constrained by the KL divergence. A distillation parameter $T$ is added to the softmax layer, and the softening degree of the soft label can be controlled by $T$. The sample size of the data set is $N$, and the number of identity tags is $K$. After the fully connected layer, the
The prediction vector of the output training sample is \( z = [z_1, z_2, z_3, \ldots z_k] \). The soft label of teacher network output is the prediction probability of \( x_i \), which is expressed as \( p_j(x_i | \theta^t) = \frac{\exp(z_i)/T}{\sum_{i=1}^K \exp(z_i)/T} \). The soft label of student network output is the prediction probability of \( x_i \), which is expressed as \( q_j(x_i | \theta^s) = \frac{\exp(z_i)/T}{\sum_{i=1}^K \exp(z_i)/T} \).

For the soft labels of teacher network and student network, KL divergence constraint loss function \( L_{kl} \) is used for distribution approximation.

\[
L_{kl} = - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K p_j(x_i | \theta^t) \log (q_j(x_i | \theta^s))
\]  

(2)

For the hard labels of student networks, the classification loss function \( L_{id} \) is used to restrict the accuracy of identity labels. Total loss function is \( L \).

\[
L_{id} = - \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K t_j \log q_j(x_i | \theta^s)
\]  

(3)

\[
L = \alpha * L_{kl} + (1 - \alpha) * L_{id}
\]  

(4)

4. Experiments and discussions

4.1. Datasets and evaluation metrics

Our main evaluation method based on Market-1501[10]. Because it is a large-scale dataset with multiple cameras and multiple persons. Market-1501 contains 32668 pictures of 1501 identities. These images were captured from 6 cameras placed in front of the supermarket in Tsinghua University. It is divided into three parts: 751 identities in 12936 images for training; 750 identities in 19732 images for testing; another 750 identities in 3368 images for querying. For simplicity, we use "Market" to represent it.

In the calculation, we use a dataset as the target domain and another dataset as the source domain. Cumulative Matching Characteristic (CMC top-1) and mean average precision (mAP) are used as evaluation indicators.

4.2. Implementation details

The experiment in this chapter is based on three platforms: server, personal computer and cloud desktop. The server’s detailed hardware environment is as follows: Inter(R) Core i7-8700K CPU, two GTX-
1080TI GPUs, 32G RAM. The server’s detailed software environment is as follows: Ubuntu 18.04, Python 3.6, PyTorch 1.1, CUDA 10.0, CUDNN 7.5. The personal computer’s detailed hardware environment is as follows: Inter(R) Core(TM) i5-8250U CPU, NVIDIA GeForce MX150 GPU, 8G RAM. The personal computer’s detailed software environment is as follows: Pycharm, Python 3.6, PyTorch 1.1, CUDA 10.0, CUDNN 7.5. The server’s detailed hardware environment is as follows: Inter(R) Core i7-8700K CPU, two GTX-1080TI GPUs, 32G RAM. The server’s detailed software environment is as follows: Ubuntu 18.04, Python 3.6, PyTorch 1.1, CUDA 10.0, CUDNN 7.5. The cloud desktop’s detailed hardware environment is as follows: Aliyun ECS 2GB. The cloud desktop’s detailed software environment is as follows: CentOS 7.3, Python 3.6, PyTorch 1.1, CUDA 10.0, CUDNN 7.5.

ResNet50 pre-trained on ImageNet is used as our teacher network. ResNet18-DSC pre-trained on ImageNet is used as our student network. To fine-tune the model by using the multi-label classification framework, fully connected layers are modified to adapt to different datasets. In order to avoid overfitting, a dropout layer is inserted before fully connected layers. The dropout rate is set to 0.5. The inputs are RGB three-channel images and are resized to 256×128. Stochastic gradient descent (SGD) momentum is 0.8. Mini-batch is 16. Epoch is 50 for the first phase and 30 for the second phase and 20 for the third phase. Experiments show that the model will converge after the 20th epoch. The initial learning rate is 0.001.

4.3. Experimental results
This experiment discussed the influence of weight coefficient α and distillation coefficient T on the experimental results of model compression. Table 1 is the influence of model compression parameters on experimental results. Compared with the baseline of ResNet18 backbone network, the method of knowledge distillation has been greatly improved on Rank-1 and mAP. When α = 0.6, T = 10, the performance of the compression model is improved by 3.1 % on Rank-1 and 5.5 % on mAP. This experiment shows that knowledge distillation can transfer the knowledge learned by high-performance large networks to low-performance small networks.

| α   | 0.6 | 0.9 |
|-----|-----|-----|
| T   | 1   | 6   | 10  | 1   | 6   | 10  |
| Rank-1 | ↑ 1.6 | ↑ 2.2 | ↑ 3.1 | ↑ 2.1 | ↑ 2.3 | ↑ 2.0 |
| mAP  | ↑ 2.4 | ↑ 3.7 | ↑ 5.5 | ↑ 3.9 | ↑ 4.2 | ↑ 3.8 |

Table 2 is a comparison before and after model compression. The number of parameters of ResNet18-DSC is only 5.6 % of that of ResNet50. The calculation amount of ResNet18-DSC is only 5.2 % of that of ResNet50. The recognition speed of ResNet18-DSC is also increased by more than ten times. Compared with ResNet50, the Rank-1 accuracy of ResNet18-DSC is only reduced by 1.6 %. Compared with ResNet50, the mAP accuracy of ResNet18-DSC is only reduced by 5.0 %. In the case of less reduction of Rank-1 accuracy and mAP accuracy, the number of parameters and computation can be reduced by nearly 20 times, and the average recognition speed can be increased by more than 10 times. Experiments show that the knowledge learned from complex networks can be transferred to a simple network by knowledge distillation.

|                      | ResNet50 | ResNet18-DSC(after model compression) | ResNet18-DSC(before model compression) |
|----------------------|----------|---------------------------------------|----------------------------------------|
| parameters (10^6)    | 22.9     | 1.3                                   | 1.3                                    |
| capacity (10^9)      | 2.87     | 0.15                                  | 0.15                                   |
| average recognition speed (s) | 1.2   | 0.1                                   | 0.1                                    |
| Rank-1(%)            | 85.6     | 84.0                                  | 75.8                                   |
| mAP(%)               | 67.8     | 62.8                                  | 52.1                                   |
ResNet18-DSC trained by knowledge distillation has a faster recognition speed in the server environment. Table 3 is comparing the average recognition speed of ResNet50 and ResNet18-DSC in server, personal computer and cloud desktop.

Table 3. Comparing the average recognition speed of ResNet50 and ResNet18-DSC in server, personal computer and cloud desktop environment.

|                  | server | Personal computer | cloud desktop |
|------------------|--------|-------------------|--------------|
| RAM (G)          | 32     | 8                 | 2            |
| CPU              | Inter(R) Core i7-8700K | Inter(R) Core(TM) i5-8250U | Inter(R) Xeon(TM) Platinum 8163 |
| GPU              | GeForce GTX-1080TI | GeForce MX150 | none         |
| ResNet50 average recognition speed (s) | 0.4 | 1.6 | - |
| ResNet18-DSC average recognition speed (s) | 0.1 | 0.5 | 1.9 |

5. Conclusions
This paper proposed a pedestrian re-identification compression method based on a depthwise separable convolution network, which effectively reduces the calculation amount of the model and improves the calculation speed of the model. In this paper, ResNet18 is used as the backbone network, in which the convolution operation is modified to the separable convolution operation, and ResNet18 (ResNet18-DSC) based on the deep separable convolution network is proposed. First, ResNet50 is used as a teacher network and ResNet18-DSC as a student network. Then, the soft label obtained by the teacher network through softmax and distillation coefficient T is used to supervise the training of the student network, and the KL divergence loss function is used to make the soft label distribution of the student network approximate the soft label distribution of the teacher network. The recognition accuracy of the student network is constrained by classification loss function. By comparing the operation effect of compression model in server, personal computer and cloud desktop environment, it provides a feasible reference for the operation of pedestrian re-identification task on devices with poor computing ability.

References
[1] Jia, M., Zhai, Y., Lu, S., Ma, S., Zhang, J. "A similarity inference metric for RGB-infrared cross-modality person re-identification". In Proc. the Twenty-Ninth International Joint Conference on Artificial Intelligence Main track. pp. 1026-1032.
[2] X. Jin, C. Lan, W. Zeng, Z. Chen and L. Zhang, "Style Normalization and Restitution for Generalizable Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3140-3149, doi: 10.1109/CVPR42600.2020.00321.
[3] Dobrowski S Z, Ustin S L, Wolpert J A. Grapevine dormant pruning weight prediction using remotely sensed data[J]. Australian Journal of Grape and Wine Research, 2003, 9(3).
[4] Ercegovac, M.D, Lang T. Redundant and On-Line CORDIC: Application to Matrix Triangularization and SVD[J]. IEEE Transactions on Computers, 1990, 39(6):725-740.
[5] Romero A, Ballas N, Kahou S E, et al. FitNets: Hints for Thin Deep Nets[C]// ICLR. 2015.
[6] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning.
[7] Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning.
[8] Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification.
[9] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2016.
[10] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 79-88, doi: 10.1109/CVPR.2018.00016.