Pedestrian Re-identity Based on ResNet Lightweight Network

Xingxing Li*, Chao Duan, Panpan Yin and Ningxing Wang

School of Engineering, Guangzhou College of Technology and Business, Foshan, 528138, China

*Corresponding author: wslxx@jxstnu.edu.cn

Abstract. With the development of deep learning technology, pedestrian re-identity technology has been widely used in multi-target tracking and cross mirror tracking tasks. In this paper, the classical deep learning ResNet18 network is used for pedestrian re-identity tasks. The advantage of the network is that it can easily realize lightweight deployment. In addition, the labeled smooth cross entropy loss function and migration learning technology are used in the process of training the network, which can realize the accuracy of map 67.8 on the Market1501 data set while lightening the network, and promote the engineering landing of pedestrian re-identity network.

Keywords: Deep learning, pedestrian re-identity, cross entropy, migration learning, ResNet.

1. Introduction

With the breakthrough of deep learning technology, artificial intelligence began to develop rapidly all over the world at an unprecedented speed. The target detection and tracking technology based on deep learning technology has been widely used in the landing of industrial projects. Especially in the traffic scene, there will be a large number of pedestrian and different types of vehicle information. These objects tend to move in real time and have different degrees of occlusion. In addition, the traffic scene will also have light and dark lights and different interference objects in different periods with the weather. If you want to capture real-time road condition information in the traffic scene, you first need to be able to adapt to various weather, time periods and other complex conditions to detect and track objects in the traffic environment in real time, and you need to use some lightweight deep learning networks to deploy and land [1].

In the real traffic scene, if you want to achieve stable tracking of pedestrian targets in the image, you need to be able to detect the target stably and track the target stably. The target in the real scene is often affected by objective factors such as image occlusion and deformation, resulting in target loss. If you want to use the tracking algorithm to find the target again, you need to use the re-identity technology. Due to the appearance representation information gap between the target and the target, for example, there are multiple pedestrians in an image, and there are also differences in representation characteristics among these pedestrians, such as clothes, height, appearance, etc. We can use the deep learning technology to make the network automatically learn and extract the characterization information. Through the comparison and matching between targets, we can get the matching similarity score of each target, and find the target again according to the ranking of the score [1].
According to the method, the design process is divided into three parts: feature learning, measurement learning and ranking optimization. Researchers' methods usually improve on these three aspects with different emphasis. Some propose novel feature learning methods, some propose effective measurement loss functions, and some optimize in the test retrieval stage. At the end of this chapter, it also summarizes the existing common data sets and evaluation indicators, as well as the analysis of the advantages and disadvantages of the existing SOTA. Feature learning methods include global feature learning: using the global image of the whole body for feature learning. Common improvement ideas include attention mechanism, multi-scale fusion and so on.

Local feature learning: use local image areas (pedestrian components or simple vertical area division) to learn features, and aggregate to generate the final pedestrian feature representation; Auxiliary feature learning: use some auxiliary information to enhance the effect of feature learning, such as semantic information (such as pedestrian attributes), perspective information (different orientation information presented by pedestrians in the image), domain information (such as the data under each camera represents a class of domain), information generated by GAN \[2\] (such as generating pedestrian images), data enhancement, etc. Video feature learning: use some video data to extract timing features and integrate multi frame image information to construct a specific network design for human feature expression: use the characteristics of Re-ID task to design some fine-grained, multi-scale and other related network structures to make them more suitable for Re-ID scenes.

Early measurement learning mainly designed different types of distance / similarity measurement matrices. In the era of deep learning, it mainly includes the design of different types of loss functions and the improvement of sampling strategies: identity loss: regard the training process of Re-ID as an image classification problem, and different pictures of the same pedestrian as a category. The common ones are softmax cross entropy loss function; Verification loss: take Re-ID training as an image matching problem. Whether it belongs to the same pedestrian for binary classification learning. Common examples include contrast loss function and binary classification loss function; Triple loss: regarding Re-ID training as an image retrieval problem, the feature distance of the same pedestrian image is less than that of different pedestrians, and its various improvements; Improvement of training strategy: adaptive sampling method (sample imbalance, difficulty) and different weight allocation strategies. After obtaining the initial search ranking results with the learned Re-ID features, the similarity relationship between images is used to optimize the initial search results, mainly including re ranking and rank fusion.

In order to achieve industrialization, this paper uses ResNet18 \[3\] network as the backbone, takes the re-identification problem as the classification problem, uses softmax \[4\] cross entropy loss function for training, uses Euclidean distance as the matching similarity measure in the process of model reasoning, and judges whether the nearest target is the same target according to the ranking of distance \[3\].

2. Principle

2.1. Deep learning ResNet18 network

There are five main forms of ResNet: ResNet18, ResNet34, ResNet50, ResNet101, ResNet152 \[3\]; In order to realize lightweight deployment, this paper tests the real-time performance of ResNet networks with different depths, and finally chooses to useResNet18 network model. As shown in Figure 1, each network includes three main parts: input part, output part and intermediate convolution part (the intermediate convolution part includes four stages from stage1 to Stage4 as shown in the figure). Although ResNet has rich variants, they all follow the above structural characteristics. The difference between networks mainly lies in the difference of block parameters and number of intermediate convolution parts.
Figure 1. Structure diagram of ResNet network model.

The forward propagation of ResNet specifies the flow direction of network data: (1) after entering the network, the data first passes through the input part (Conv1, BN1, Relu, Max-pool); (2) Then enter the intermediate convolution part (layer1, layer2, layer3, layer4, where the layer corresponds to the stage we mentioned earlier); (3) Finally, the results are obtained through an average pool and full connection (FC) output; Specifically, the differences between ResNet18 and other res series networks are mainly in layer1-layer4, and other components are similar. For the network input part, all ResNet network input parts are composed of a large volume core with size = 7×7 and stripe = 2 and a maximum pool with size = 3×3 and stripe = 2. Through this step, a 224×224 input image will change into a 56×56 feature image, greatly reducing the size required for storage; For the network input part, all ResNet network input parts are composed of a large volume core with size = 7×7 and stripe = 2, and a maximum pool with size = 3×3 and stripe = 2. Through this step, a 224×224 input image will become a 56×56 feature image, greatly reducing the size required for storage. The residual block implements the basic block shown in the figure below. The input data is divided into two paths. One path is convoluted by two 3×3 convolutions, and the other path is directly shorted. The two are added and output through relu.

Figure 2. Structure diagram of residual network.
The network output part is very simple. Through global adaptive smoothing pooling, all feature maps are pulled into 1×1. For ResNet18, the input data of 1×512×7×7 is pulled into 1×512×1×1, and then connected to the full connection layer output. The number of output nodes is consistent with the number of prediction categories. In the model prediction, the last full connection layer needs to be removed and turned into a 1×512 dimensional fixed feature vector for feature comparison.

2.2. Label smoothing cross entropy loss function
Label smoothing, also known as label smoothing, is actually a regularization method to prevent overfitting. The traditional classification loss adopts softmax loss. Firstly, softmax is calculated for the output of the full connection layer, which is regarded as the confidence probability of each category, and then the loss is calculated by cross entropy. The formula is as follows:

\[
q_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}
\]

\[
\ell(p, q) = -\sum_{i=1}^{K} p_i \log(q_i)
\]

In this process, the output probability of each sample in the correct category should be 1 as much as possible, so that the corresponding Z value is +∞, which increases the distance between it and other categories. Now suppose that the label of a multi classification task is [1, 0, 0]. If there is a problem with its own label, it will do great harm to the model, because a non-class sample is forced to learn in the process of training, and its probability is very high, which will affect the estimation of a posteriori probability. And sometimes the between classes are not unrelated. If the difference between the probability of encouraging output is too large, it will lead to over fitting to a certain extent. Therefore, the idea of label smoothing is to change the target from one hot label to the following form:

\[
q_i = \begin{cases} 
1 - \varepsilon & \text{if } i = y \\
\frac{\varepsilon}{K-1} & \text{otherwise}
\end{cases}
\]

Where \(\varepsilon\) is a small constant, which makes the probability optimal objectives in softmax loss no longer be 1 and 0, and the optimal solution of Z value is no longer positive infinity, but a specific value. This avoids over fitting to a certain extent and alleviates the impact of wrong labels.

2.3. Transfer learning
This paper uses the transfer learning technology when using ResNet18 network training data. During training, only the feature extraction layer is opened in the first few cycles. Let the feature extraction layer learn for several cycles, and then solidify the feature extraction layer. In the next training process, there is no iteration, and the last full connection layer is opened for separate training. The optimizer selects Adam, and the learning rate strategy is used to adjust the learning rate at equal intervals.

3. Experiments
The data used in this experiment is the market-1501 data set, which was collected on the campus of Tsinghua University, photographed in summer, constructed and published in 2015. It includes 1501 pedestrians captured by 6 cameras (including 5 high-definition cameras and 1 low-definition camera) and 32668 detected pedestrian rectangular boxes. Each pedestrian is captured by at least 2 cameras and may have multiple images in one camera. The training set has 751 people, including 12936 images, with an average of 17.2 training data per person; There are 750 people in the test set, including 19732 images, with an average of 26.3 test data per person. The pedestrian detection rectangle of 3368 query images is drawn manually, while the pedestrian detection rectangle in the gallery is detected by DPM detector. The fixed number of training sets and test sets provided by the data set can be used under single shot or multi shot test settings. The re-identity effect is shown in Figure 3:
Figure 3. Schematic diagram of segmentation results of dirty silicon image.

The leftmost pedestrian in the figure is the target to be retrieved, and the corresponding group of images on the right are the retrieval results in the image retrieval library. The green box represents the correctly retrieved target, and the red box represents the correct target that has not been recognized. It is sorted from high to low according to the confidence. In this group of figures, a total of 10 retrieval results represent the ranking of Rank1 to rank10. From the re-identity effect, the target image only leaves the back, and the algorithm can match the corresponding target in a large number of retrieval libraries.

The accuracy of re-identity is counted, as shown in Table 1 below:

| Datasets   | Map  | Rank1 | Rank5 | Rank10 |
|------------|------|-------|-------|--------|
| Market1501 | 67.8%| 84.1% | 93.3% | 95.7%  |

As can be seen from Table 1, the overall accuracy of map using lightweight network ResNet18 can reach 67.8%, of which Rank1 is 84.1%. In the real scene, the tracked target has no longer time span than the data set Market1501, and the recognition result is often better. This experiment fully proves that ResNet18 network can not only realize lightweight deployment, but also balance accuracy, which is more suitable for deployment on some edge embedded devices.

4. Conclusion
In this paper, the classical deep learning ResNet18 network is used for pedestrian re-identity task. The advantage of the network is that it can easily realize lightweight deployment. In the process of training the network, the labeled smooth cross entropy loss function and migration learning technology are used, which can realize the lightweight network and the accuracy of map 67.8% on the Market1501 data set. It is suitable to deploy the network to edge computing devices such as embedded devices.

Acknowledgments
This work was financially supported by fund project, that is, Young Talents in Higher Education of Guangdong, China, (No. 2021KQNCX125 and No. 2019KQNCX232).

References
[1] Duan C, Li X. Multi-target Tracking Based on Deep Sort in Traffic Scene [J]. Journal of
Physics: Conference Series, 2021, 1952 (2): 022074 (6pp).

[2] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks [J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[4] Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax [J]. arXiv e-prints, 2016.