Target detection algorithm based on improved multi-scale SSD

Jianying Cao1*, Yan Kong2, Xinlu Zhang3, Yongjia Li4 and Xiaofeng Xie5

1 School of information engineering, Longdong university, Qingyang, Gansu, 745000, China
2School of Automation, Central South University, Changsha, Hunan, 410000, China
3School of Information Sciences and Technology, Northwest University of Chang’an Campus, Xi’an, Shaanxi, 710000, China
4Peninsula Grammar School, Melbourne, Victoria, 3000, Australia
5College of Information Engineering, Wuchang Institute of Technology, Wuhan, Hubei 430065, China.

*Corresponding author’s e-mail: cao_jianying@163.com

Abstract. The traditional SSD algorithm has a serious abstraction of feature extraction content, which makes it difficult to achieve effective detection of small targets. At the same time, the problem of feature layer fusion is difficult due to different scales. In this paper, an improved SSD based target detection algorithm is proposed. By introducing feature enhancement method, the adjustment steps of high-level feature size are omitted, Which makes it unnecessary to reduce the dimension of features, and at the same time, it uses the multi-scale candidate area which accords with the proportion of pedestrians in the detection network to enhance the feature extraction ability of small targets, effectively improves the accuracy and operation speed of SSD algorithm, and saves the loss of the network.

1. Introduction

Target detection is widely used in image analysis, assistant driving, intelligent video monitoring and other fields [1]. The natural environment has the characteristics of changeable light and large interference. The target detected in the video image has the problems of small size and scale change. At present, convolutional neural network can replace the traditional manual design features, which has stronger feature expression ability and better robustness [2], and has made a lot of progress in the field of target detection, before and after the emergence of many neural network-based detection algorithms. Girshick proposed Fast R-CNN algorithm based on RCNN algorithm, which greatly improved the efficiency and accuracy of the algorithm. However, fast r-cnn still hasn't solved the problem of slow selection of candidate regions. In order to better solve this situation, the Fast RCNN algorithm proposed by Ren revealed the concept of region generation network, which was used to select candidate regions faster and achieved good results. The SSD and YOLOv2 algorithms proposed by Liu and Joseph achieve real-time target detection and ensure high detection accuracy [3-6].
2. Traditional SSD algorithm
SSD algorithm is a kind of target detection algorithm, which directly convolutes the whole image and uses multi-scale feature layer of pyramid structure to predict the coordinates and categories of target bounding box. In the feature layer of pyramid structure, the size of the shallow feature map is more obvious than the large detail feature, which is more suitable for detecting small targets, and the deep feature map has strong semantics for detecting large targets. The SSD algorithm takes VGG16 as the main network, and the convolution feature layer added after VGG16 is also a structure of decreasing size layer by layer, which is used to generate detection and prediction values of different scales. SSD algorithm evaluates and predicts the default frame of convolution output with different aspect ratio. Each grid on the feature map will have a series of fixed size boxes called default boxes. Each grid has K default boxes, each of which predicts C target categories and 4 offsets. In the pyramid feature layer of the original SSD algorithm, the feature information of different scale features lacks each other's feature information. Because of the large receptive field in the high level feature layer, the content of feature extraction is too abstract, which is not conducive to the detection of small targets. These problems make the original SSD algorithm have limitations in dealing with small target detection tasks. The traditional SSD frame is shown in Figure 1.

3. Improved SSD target detection algorithm
3.1 Feature enhancement strategy
In this paper, a feature enhanced SSD algorithm is proposed. In order to omit the feature dimension reduction operation after feature fusion and reduce the overall loss of the network, firstly, the convolution operations of Conv8, Conv7, Conv6, Conv4, FC7, which are different scale feature layers, are checked by 3×3 convolution with padding=1, and the convolution dimension is required to be no higher than the original feature dimension. After convolution, five different new feature layers are obtained, and the network framework is shown in Figure 2. Compared with the original feature layer, the semantic features of each feature layer are more obvious. And because the edge is supplemented by the convolution process with constant feature size, the bottom details and edge features in the feature map are better preserved. In the convolution feature scale of each layer, the algorithm in this paper is also larger than the original one. In the process of feature fusion, the size of the feature image does not need to be adjusted, and the edge features of the feature image are not lost. In order to match the shallow features with the deep features, the original feature fusion strategy needs to put the deep features into bilinear interpolation, which has high error and more network loss. In this paper, the deep features do not need to be fine-tuning the size of the lightweight feature fusion method, the fused features do not need to reduce the dimension operation and calculation, greatly improving the detection efficiency of the whole network, reducing the loss. To sum up, the multi-scale SSD algorithm based on feature enhancement is faster than the existing SSD algorithm, with only 2.5 frames/s lost.
3.2 Multi-scale pedestrian detection network

In order to solve the problem that SSD shallow network has poor ability to extract small target features and the horizontal to vertical ratio of some candidate areas in the existing SSD algorithm detection network is not suitable for pedestrians. We set the horizontal to vertical ratio of candidate regions in SSD algorithm to 1:1, 1:2 and 1:3. Select convolution layer 1, layer 2, layer 3, layer 4 and eigengraph fusion network as the output part of our overall network, and the structure framework is shown in Figure 3.

\[
S = \frac{|A \cap B|}{|A \cup B|} \quad (1)
\]

In this formula, \(S\) is the similarity, \(A\) is the real region of the target, and \(B\) is the range of candidate regions. The policy is set as follows: when the ratio of the candidate region to the real region is \(S > 60\%\), the real target exists in the candidate region, otherwise there is no target to be detected in the target region.

In order to extract the corresponding regional features, we need to pool the region of interest (ROI) with the detection network, and finally use the classification of candidate frame and regression network to generate the prediction and classification of the target pedestrian. Border return network output is the border position parameter

\[
T^k = (T_x^k, T_y^k, T_w^k, T_h^k) \quad (2)
\]

In formula (2), \(k\) represents the category. The loss of border regression layer is:

\[
L_{reg}(T^u, V) = \sum smooth_{k} (T^u_j - V_j) \quad (3)
\]

In formula (3),
The classification and regression network of candidate frames are trained by joint loss function, that is

\[ L(p, u, T^u, V) = L_{cls}(p, u) + \lambda[u \geq 1] \cdot L_{reg}(T^u, V) \]  

(5)

Where \( L_{cls}(p, u) = -\log(p_u) \) is the log loss of the real class \( u \). \( L_{reg} \) is activated only if the area to be tested is target, that is \( p^*_u = 1 \).

4. Experiment and results

4.1 Data set

In this paper, Pascal VOC 2012 test set is selected as the experimental data set, which has 35845 images. In this paper, the test set images are divided into four categories, including 7979 simple background single target images, 7946 simple background multi-target images, 9871 complex background single target images, and 10049 complex background multi-target images.

4.2 Model training

In this paper, Tensorflow framework is chosen as the experimental platform of deep learning. According to the current standard strategy of target detection method based on deep learning, the training network of this paper is initialized by the model that has been pre-trained in the task of Imagenet classification. The weight of each volume layer in the feature extraction network is initialized by the VGG16 network trained by Imagenet. SGD is used to optimize the whole training network model. The experimental environment is python2.8, GeForce GTX 1660 Ti.

4.3 Results and analysis

In the experimental part of this paper, the multi-scale SSD pedestrian detection algorithm based on feature enhancement is compared with the existing SSD algorithm and similar pedestrian detection algorithm. All algorithms are implemented in VOC 2012 data set. The performance difference between this algorithm and the existing SSD algorithm is evaluated by two indexes: center position error and coverage. Center position error refers to the Euclidean distance between the real center position of the target in the original image and the center position of the candidate frame located by the detection network. The smaller the center position error is, the higher the positioning accuracy is. The comparison results are shown in Table 1, where TPA represents the algorithm proposed in this paper.

| Image type                     | SSD | TPA |
|--------------------------------|-----|-----|
| Simple background single target | 4.8 | 2.1 |
| Simple background multi-target  | 17.6| 8.4 |
| Complex background single target| 25.2| 14.3|
| Complex background multi-target | 42.4| 19.6|

Coverage rate refers to the proportion of the overlapped part of the real target and the detection network location frame in the total range of the image. The expression of coverage rate is

\[ C = \frac{|S_T \cap S_G|}{|S_T \cup S_G|} \]  

(6)

\( S_T \) represents the real location range of the target in the image, \( S_G \) represents the location range of the target location frame. The larger the coverage, the higher the target location accuracy. The comparison results of coverage indicators are shown in Table 2.
Table 2. Comparison of coverage rate of VOC 2012

| Image type               | SSD  | TPA  |
|--------------------------|------|------|
| Simple background single target | 87.2 | 95.1 |
| Simple background multi-target | 77.5 | 88.7 |
| Complex background single target | 81.2 | 91.2 |
| Complex background multi-target | 71.3 | 85.9 |

According to the experimental results in Table 1 and Table 2, the improved SSD algorithm has lower center position error and better coverage in most images of VOC2012 data set. Compared with the existing SSD algorithm, it also shows the superiority of this algorithm. In the VOC2012 data set, this paper compares the improved SSD algorithm with the existing mainstream target detection algorithms such as SSD algorithm, RCNN algorithm, YOLOv2 algorithm and R-FCN algorithm. The comparison results are shown in Table 3.

Table 3. Comparison results of detection algorithms

| Algorithm | Basic network | Pre-trained | Map/% | FPS(frame/s) |
|-----------|---------------|-------------|-------|--------------|
| SSD       | VGGNet        | √           | 75.4  | 21.0         |
| RCNN      | VGGNet        | √           | 70.1  | 5.2          |
| YOLOv2    | DarkNet-19    | √           | 72.8  | 42.3         |
| R-FCN     | ResNet-50     | √           | 81.3  | 8.6          |
| TMA       | VGGNet        | √           | 80.5  | 31.4         |

It can be seen from Table 3 that on the map index, the algorithm proposed in this paper is 5.1% higher than the existing SSD algorithm, 10.4% higher than the RCNN algorithm, but 0.8% lower than the R-FCN algorithm, because the basic network of the r-fcn algorithm is ResNet, which has a deeper network and better performance than VGG16. However, the disadvantage of R-FCN algorithm is that it adopts a more complex U-shaped network structure, and there is a waiting time before the feature fusion step. These problems cause that the r-fcn model is too complex in structure and the loss is very large. Compared with the 8.6 frame/s of F-RCN FPS, the detection speed of this algorithm can reach 31.4 frame/s, which is much faster than the F-RCN algorithm.

5. Conclusion

This paper combines the idea of feature enhancement and multi-scale feature fusion to improve the original network structure of SSD algorithm. In view of the poor feature extraction ability of the existing SSD shallow network for small targets and the fact that the horizontal to vertical ratio of some candidate areas in the existing SSD algorithm detection network is not suitable for small targets, we adopt the feature enhancement strategy and multi-scale candidate areas more suitable for pedestrian proportion. We compare the algorithm with other mainstream target detection algorithms in VOC2012 data set, and the algorithm in this paper is obviously superior to the existing SSD algorithm in the two indicators of center position error rate and coverage rate. Compared with other mainstream target detection algorithms, this algorithm has some advantages in map and fps.

References

[1] Samson, C., Blanc-Féraud, L., Aubert,G., et al. (2000) A Level Set Model for Image Classification. International Journal of Computer Vision, 40:187-197.
[2] Moeskops, P., Viergever, M.A., Mendrik, A.M., et al. (2016) Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. IEEE Transactions on Medical Imaging, 35:1252-1261.
[3] Zhang, L., Liu, J., Zhang, B., et al. (2019) Deep Cascade Model-Based Face Recognition: When Deep-Layered Learning Meets Small Data. IEEE Transactions on Image Processing., 29: 1016–1029.
[4] Lu, J.W., Liong, V.E, Wang, G., et al. (2015) Joint Feature Learning for Face Recognition. IEEE Transactions on Information Forensics and Security., 10: 1317–1383.
[5] Gao, S.H., Zhang, Y.T., Jia, K., et al. (2015) Single Sample Face Recognition via Learning Deep Supervised Autoencoders. IEEE Transactions on Information Forensics and Security., 10: 2108–2118.

[6] Lu, J.W., Wang, G., Zhou, J. (2017) Simultaneous Feature and Dictionary Learning for Image Set Based Face Recognition. IEEE Transactions on Image Processing., 26: 4042–4054.