Research on robot target recognition based on deep learning

Zhenyu Sun¹, Xiaoming Guo², *, Xiaoyang Zhang³, Jiangxue Han⁴ and Jian Hou⁵

¹,²,³,⁴,⁵Computer and Communication Engineering College, Liaoning Shihua University, Fushun, China

*Corresponding author email: panbin@lnpu.edu.cn

Abstract: For the traditional machine vision recognition technology in the industrial field can not handle the problem of different classes of workpieces placed randomly and stacked on each other, this paper improves the SSD algorithm model based on the research of deep learning target detection algorithm. Firstly, a depth-separable convolutional structure is introduced to optimize the VGG backbone feature network. Then a multi-level feature fusion mechanism is introduced in the prediction part to increase the semantic information of features. Qualitative and quantitative experimental results show that the improved optimization method of the SSD model in this paper is validated well on the dataset, and the improved SSD model mAP value is increased by 4.3% compared with the original, and the detection speed is increased by nearly two times, thus proving the effectiveness of the improved method.

1. Introduction

The current machine vision technology also has some drawbacks and limitations. In the common industrial production line, the workpieces are often placed randomly and stacked on each other, and the traditional robot vision technology can no longer meet the actual demand[1]. Mainly limited to the artificial design of the algorithm, the traditional method needs to identify and grasp the workpiece before the use of mechanical methods to first place the target workpiece in a certain order unobstructed on the plane, and then according to the algorithm designed in advance to drive the robot to complete a series of movement tasks, which wastes time and reduces the efficiency to a certain extent [2]. In actual industrial production, a large number of workpieces are stacked randomly and scattered, and there are overlaps and occlusions between workpieces[3]. At this time, the traditional machine vision technology can not meet the actual needs, so in the actual production of different workpiece stacking recognition detection has an urgent need, but also has a high research value.

Target detection based on deep learning has been a research hotspot since its inception, especially deep learning in the convolutional neural network and the related engineering technology is also to promote the development of continuous optimization and improvement. In recent years, CVPR, ECCV and other top international computer vision conferences continue to emerge new target detection model algorithms, the emergence of these high-quality model algorithms have improved the speed and accuracy of target detection to varying degrees, and some networks have been able to achieve real-time processing capabilities. Therefore, the integration of deep learning technology into industrial robot vision target detection will definitely improve the level of automation and intelligence of industrial production substantially, and reduce the cost and resource consumption to improve production efficiency.

Based on the above background, this paper takes various industrial components as an example, conducts profound research on target recognition detection of robots, applies the relevant theoretical techniques of deep learning[4], including image processing operations and target detection algorithms,
to the robot recognition system model, and improves the overall model performance according to the relevant optimization algorithms, so that the robot can perform target detection more quickly and improve the robot's intelligent target recognition detection and recognition capability.

2. Improved algorithms

2.1. Depthwise Separable Convolution
Depthwise Separable Convolution (DSC) can be understood as an efficient decomposition of the standard form of convolution, and was first proposed in the paper MobileNet [5] for lightweight networks. The principle is to decompose the standard form of convolution into depthwise convolution and 1×1 point convolution, which can largely reduce the number of parameters and operation cost compared to the conventional convolution operation, where depthwise convolution is responsible for filtering feature learning, while 1×1 point convolution is responsible for reorganizing the filtered features.

As shown in Figure 1 is the computational flow of the depth-separable convolution, the input feature size is $(F_D, F_D, M)$, the size of the convolution kernel is $(D_k, D_k, M, N)$, and the size of the output feature map is $(G_D, G_D, N)$, which is calculated by the following equation.

$$G_{k,j,n} = \sum_{i,j,m,n} K_{i,j,m,n} F_{k+i-1,1+j-1,m}$$

(1)

When the input feature map size is $D_F$, then the computational size of the standard convolution is as follows.

$$C = D_K * D_K * M * N * D_F * D_F$$

(2)

When deep convolution is used, only one convolution kernel is assigned to each channel, and then the output is combined in the form of point convolution, and the depth convolution is calculated as follows.

$$G_{k,j,n} = \sum_{i,j,m,n} K_{i,j,m,n} F_{k+i-1,1+j-1,m}$$

(3)

Deep convolution, because there is only one convolution kernel per channel, has the following computational size.

$$C = D_K * D_K * M * N * D_F * D_F$$

(4)
The input channel is $M$, and the output channel is $N$, and the computational size of the $1 \times 1$ point convolution is as follows.

$$C = M \times N \times D_F \times D_F$$  \hspace{1cm} (5)

Combining the above computational steps, the combination of the depth convolution and the introduced $1 \times 1$ point convolution becomes a depth-separable convolution, and the overall computational effort is as follows.

$$C = D_K \times D_K \times M \times N \times D_F \times D_F + M \times N \times D_F \times D_F$$ \hspace{1cm} (6)

It can be easily seen that the depth-separable convolution operation is very simple and efficient, using only a single convolution kernel for the processing of the feature map and then combining the features by point-by-point convolution, which reduces the number of parameters of the network to a large extent.

This is shown in the lower figure of Figure 2, the VGG network structure before optimization stacked $3 \times 3$ convolution modules, as shown above in Figure 2 after optimization using a stacked $3 \times 3$ depth convolution and $1 \times 1$ point convolution of depth-separable convolution, depth-separable convolution can greatly reduce the number of parameters, thus reducing the amount of network computation, and also more efficient feature due to the introduction of more nonlinear. The depth-separable convolution can greatly reduce the number of parameters and thus the computational effort of the network.

![Figure 2 The improved convolution module](image)

2.2. Multi-level feature fusion

Multi-level features refer to the use of different feature maps for target detection, the difference between these feature maps is reflected in the different resolutions, multi-level features are also included in the feature pyramid of a class[7]. In the actual scene of the target detection recognition task, the recognition of large targets and small targets is very different, mainly in the resolution, smaller targets itself imaging will be lower resolution, contains less limited information, so the detection effect will be very poor, that is, the detection accuracy is very low.

The SSD model itself has been improved for image pyramids, with a total of six feature maps from which features are extracted and then output[8]. But this still has some problems, among which the shallow feature map has the advantage of being able to locate well because less location information and details are lost in the deepening of network propagation, but the semantic information carried by itself will be sparse because of the low resolution, on the contrary, the feature map of the high level is characterized by more abstraction and has rich speech category information, but a lot of location information is lost, which means that locating the target is poorer. Therefore, this multi-level feature recognition algorithm is generally effective in recognizing small objects, and since the object of this paper is some industrial components of small size, the multi-level feature fusion method is introduced for the detection of small size objects.

The multilevel feature fusion mechanism was proposed in the FPN [9] paper, and the overall principle is well understood, which is to fuse the features of different layers into one layer, so that it has the advantage of rich semantic information in the deep layer and also has the feature of good localization in the shallow layer.
As shown in Figure 3, the structure of the multi-level feature fusion network after its introduction is illustrated. Multi-level feature fusion also has many points of attention when applied, it needs to ensure that the size is consistent across different feature channels as well as resolutions[10]. The way to achieve this is to use 1×1 point convolution for channel reduction to 64 when collecting the feature map, which can meet the objective requirements of feature fusion and reduce the computational effort. In addition, a special operation is performed for each deep feature map, using bilinear interpolation for double upsampling to the same resolution of the preceding neighboring feature maps, then summing with them, and then using 3 × 3 convolution for output. Finally in the later prediction part, all the layers in the multi-level feature fusion pyramid share the prediction layer.

3. Experimental results and analysis

3.1. Comparative analysis of main network improvement experiments

Where SSD is the result of the unimproved model and S-SSD is the result of the model after improving the VGG backbone network to a deeply separable convolutional module, as shown in Table 1.

| Detection algorithm | Average accuracy value/% | Detection speed/(frame/s) |
|---------------------|--------------------------|----------------------------|
| SSD                 | 93.95                    | 4.1                        |
| S-SSD               | 93.97                    | 8                          |

The experimental results show that, firstly, the mAP value of the improved S-SSD model has exceeded the original SSD model in terms of accuracy comparison. Secondly, the CPU test S-SSD is 15 frames/s based on this paper, and the magnitude is also much higher than the original SSD model. the data results of S-SSD show the feasibility of improving the backbone network based on deep separable convolution in this paper.

3.2. Comparative analysis of results

This experiment uses 5000 industrial original datasets, of which 4500 are the training set and 500 images are the test set, and the improved SSD training loss with the original loss is shown in Figure 5.
Figure 4 loss comparison chart

From the curve in the figure 4, it can be analyzed that the total loss of the improved SSD model decreases much faster than that of the original SSD model from the beginning, and at the end when the iteration reaches 60,000 steps, the loss convergence value of the improved SSD model is about 0.3, and that of the original SSD model is about 0.6, which is nearly double. It can be seen that the improved SSD network model structure is more suitable for the identification and localization of industrial components.

4. Conclusion
The speed improvement aspect is the introduction of deep separable convolution for optimization, which reduces the parameters for network training and speeds up the network computation. For accuracy, a multi-level feature fusion mechanism is introduced in the prediction part. The final prediction feature map not only has the advantage of accurate localization at the shallow level, but also has rich semantic information at the deep level, which improves the recognition accuracy.

Acknowledgments
The National Natural Science Foundation of China under Grant Nos. 61602228, 61572290; the Revitalization Talents Program of Liaoning Province under Grant No. XLYC1807266.

References
[1] Jiang Xinsong. The development of foreign robots and our countermeasure research [J]. Robotics, 1987, 9(1): 3-10.
[2] White paper on industrial machine vision industry development in China [J]. Robotics Industry, 2020(06):76-95.
[3] Dawei Fang. The current situation and development trend of industrial robots in China[J]. Electronic Technology and Software Engineering, 2019(18):100-101.
[4] Li Y. Research on face expression recognition algorithm based on deep learning. Beijing University of Technology [D]. Beijing: Beijing University of Chemical Technology, 2017.
[5] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.
[6] Chen Keqi, Zhu Zhiliang, Deng Xiaoming, Ma Cuixia, Wang Hongan. A review of deep learning research on multi-scale target detection[J]. Journal of Software, 2021, 32(04):1201-1227.
[7] Li, Chenghao, Zhang, Jing, Hu, Li, Xiao, Xianpeng, Zhang, Hua. A small target detection algorithm based on multi-scale perceptual field fusion [J/OL]. Computer Engineering and Applications: 1-7 [2021-04-21].
[8] W. Liu, D. Anguelov, D. Erhan, et al. SSD: Single Shot MultiBox Detector[J]. In: European Conference on Computer Vision, 2016, arXiv:1512.02325.

[9] K. He, X. Zhang, S Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. In: European conference on computer vision. Springer, 2014, pp. 346–361.

[10] Liu Jianfeng, Pan Chen. Enhanced feature pyramid structure for salient target detection algorithm [J/OL]. Computer Engineering and Applications:1-12 [2021-04-21].