International Journal of Engineering Continuity

Mobile Robot Object Detection Method Based on Deep Learning

Guo Yuhan
School of Mechanical and Electrical Engineering and Automation,
Shanghai University, Shanghai, China

Abstract: The task of object detection is to accurately and efficiently identify and locate a large number of predefined categories of object instances from images. With the wide application of deep learning, the accuracy and efficiency of target detection have been greatly improved. However, deep learning-based target detection still faces challenges from key technologies such as improving and optimizing the performance of mainstream target detection algorithms, improving the detection accuracy of small target objects, realizing multi-class object detection and lightweight detection model. In response to the above challenges, based on extensive literature research, this paper analyses methods for lightweight detection models and improved detection accuracy from the perspective of YOLOv5s network structure. The problems to be solved in target detection and the future research direction are predicted and prospected.

Keywords: Target detection, YOLOv5, deep learning, mobile robot.

Introduction

The perception of external information of mobile robots is mainly based on the vision system, in which the object detection technology of mobile robots is an important means for them to understand the environment, that is, the robot understands the complex surrounding environment through the deployed sensors, recognizes all the surrounding targets, and locates them (Suwoyo, Thong, et al., 2022). Target detection based on mobile robot vision system is one of the necessary skills for mobile robot to complete tasks (Suwoyo, Hidayat, et al., 2022), and it is also a difficult problem to be solved (Suwoyo & Harris Kristanto, 2022). Traditional target detection generally uses the frame of sliding window, which mainly includes three steps: firstly, some candidate regions are selected on the image, then feature extraction is carried out

Correspondents Author:
Guo Yuhan, School of Mechanical and Electrical Engineering and Automation, Shanghai University, Shanghai, China
Email : guoyuhan01@gmail.com

Received: June 21, 2023; Accepted: July 11, 2023; Publication: July 26, 2023.

on these candidate regions, and finally, a trained classifier is used for classification (Suwoyo, Abdurohman, et al., 2022). In recent years, target detection methods based on deep learning have greatly improved the accuracy of image classification and become the mainstream algorithms in the current target detection field, such as R-CNN, Fast R-CNN and Faster R-CNN (Ostanin et al., 2022; Petrović et al., 2022). However, the detection accuracy often depends on the complex frame and high strength hardware accelerator, which cannot be directly transplanted to the robot platform. Lightweight networks such as YOLO and SSD have relatively low hardware requirements and fast detection speeds, with processing speeds up to 45 fps (Raikwar et al., 2022). The disadvantage is that the detection effect of small objects is not very good, the prediction accuracy of the frame is not very high, and the overall prediction accuracy is slightly lower than that of Fast-R-CNN.

Object detection, as one of the most fundamental and challenging problems in the field of computer vision, has been widely studied and explored in recent years. As an important task in the field of computer vision, object detection usually needs to provide the specific location of certain visual objects in digital images (Raudmäe et al., 2023). In addition, object detection is also an important part of many other tasks, such as instance segmentation, image description generation, object tracking. In the past 20 years, the development of object detection has roughly experienced two historical periods: the traditional object detection period (before 2014) and the deep learning-based detection period. The traditional target detection algorithm can be summarized as follows: First, the whole image is traversed by sliding window to generate a certain number of candidate boxes; Secondly, the features of the candidate frame are extracted. Finally, support vector machine (SVM) and other classification methods were used to classify the extracted features, and then the results were obtained (Shi et al., 2023; Sun et al., 2023).

In the early target detection task, the main way to extract features is manual extraction, which has certain limitations and the performance of manual features tends to be saturated. Target detection algorithms can be divided into two categories according to different detection ideas: two-stage detection and one-stage detection. The two-stage detection algorithm is based on the proposed candidate box (Tan et al., 2023). Firstly, the regional candidate boxes are generated, then the features of each candidate boxes are extracted, and finally the position boxes are generated and the corresponding categories are predicted, which is characterized by high precision but slow speed. The one-stage detection network is characterized by high speed but low precision when it generates candidate frames for classification and boundary frame regression (Yang et al., 2023; Yu et al., 2023).

Research Method

Convolutional neural network

Convolutional Neural Networks (CNN), a deep neural network with convolutional structure, is one of the most representative neural networks in deep learning techniques. The convolutional structure can reduce the amount of memory occupied by the deep network, reduce the number of parameters of the network and alleviate the overfitting problem of the model. It is a feedforward neural network, which is widely used in computer vision. Convolution neural network composed of input layer, hidden layer and output layer, implicit layer containing convolution, pooling layer and link layer. The input layer and output layer are used for the input of data and the output of results. The roles of convolutional layer, pooling layer and fully connected layer in convolutional neural network are described in detail in the following paper (Zhang et al., 2023).

Convolutional Layer

Convolutional layer is the core part of convolutional neural network, which is mainly used for feature extraction of input images. The convolution layer is obtained through the convolution operation (Zhang et al., 2022). The convolution operation is summed by multiplying the convolution kernel with the data of the corresponding position on the image or feature map, as the feature of the corresponding position of the next layer (Zhao & Cheah, 2023). This is done by sliding the window from the upper left corner of the image or feature map to the lower right corner. The operation process of convolution operation is shown in Figure 1.

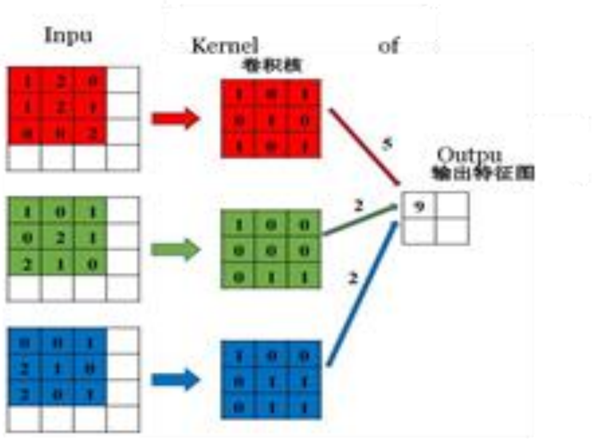
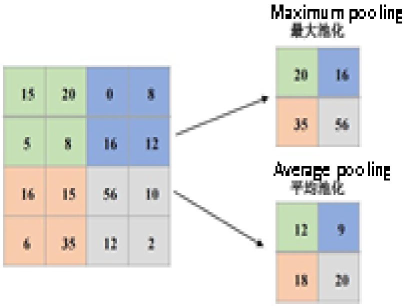


Figure 1 Schematic diagram of convolution process

It can be seen from Fig. 1 that the operation process of standard convolution is as follows: the convolution kernel slides on the input feature graph with a specific step length, and the operation is carried out by multiplying and summing corresponding window positions.

Pooling

Pooling, also known as subsampling, is used after the convolution operation to reduce the number of parameters in the network and thus speed up the training. In addition, pooling layer can also prevent the network from overfitting. Pooling generally includes maximum pooling and average pooling. The operation diagram of the two pooling modes is shown in Figure 2.



15	20	0	8
5	8	16	12
16	15	56	10
6	35	12	2

20	16
35	56

12	9
18	20

Figure 2 Schematic diagram of convolution process

As can be seen from Fig.2, the process of maximum pooling operation is as follows: select the area of the corresponding window in the input feature map, and take the maximum number of the cell values in this area as the value of the window area cell. The average pooling operation process is as follows: select the area of the corresponding window in the input feature map, add the values of all cells in the area and calculate the average result as the value of the window area cell.

Fully connected layer

The full-connection layer usually appears in the second half of the network and converts the previous multidimensional feature graph into a feature vector with fixed length, which is convenient to input into subsequent classifiers for classification. Each node of the fully connected layer is connected to each node of the previous layer, so the connection layer occupies most of the parameters of the network. In order to reduce the number of network parameters, some detection models design the structure without the full connection layer. For example, GoogLeNet uses global average pooling to replace the full connection layer, which not only reduces the number of parameters, but also makes the network structure more flexible, without limiting the input resolution of the image. However, the full connection layer is still widely used because of its simple implementation and easy to understand.

YOLOv5s algorithm

YOLOv5 is a new-generation target detection network of YOLO series, which integrates various optimization model methods based on YOLOv3 and YOLOv4. Compared with YOLOv3 and YOLOv4, YOLOv5 has the characteristics of smaller model files, shorter training time and faster reasoning speed while maintaining approximately equal detection accuracy. According to the depth of the network and the width of the feature map, YOLOv5 can be divided into four models: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Among them, YOLOv5s has the fastest processing speed and YOLOv5x has the highest detection accuracy. YOLOv5s network structure is composed of Input, Backbone, Neck and Head, as shown in Figure 3.

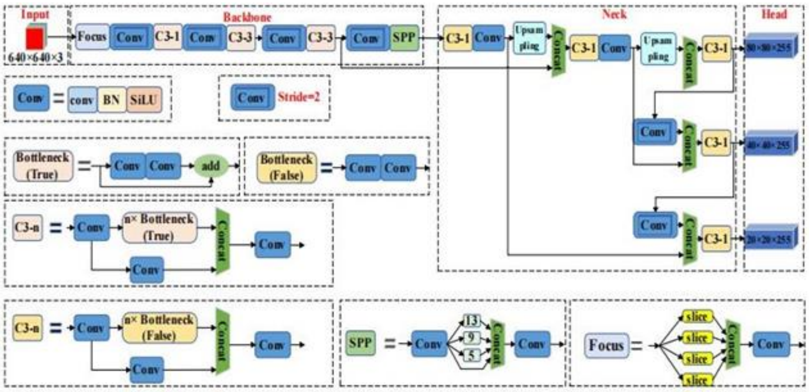

Input
640×640×3
Backbone
Focus
Conv
C3-1
Conv
C3-3
Conv
C3-3
Conv
SPP
Neck
C3-1
Conv
Upsampling
Concat
C3-1
Conv
Upsampling
Concat
C3-1
Conv
Concat
C3-1
Conv
Concat
C3-1
Head
80×80×255
40×40×255
20×20×255
Conv = conv BN SiLU
Conv Stride=2
Bottleneck (True) = Conv Conv add
Bottleneck (False) = Conv Conv
C3-n = Conv n× Bottleneck (True) Conv Concat Conv
C3-n = Conv n× Bottleneck (False) Conv Concat Conv
SPP = Conv 13 9 5 Concat Conv
Focus = slice slice slice slice Concat Conv

Figure 3 YOLOv5s network structure

The input mainly includes Mosaic data enhancement, image size processing and adaptive anchor frame calculation. Mosaic data enhancement combines four images to enrich the image background. Image size processing adaptively adds the least black edge to the original image of different length and width, and uniformly scales to the standard size. Adaptive anchor frame calculation Based on the initial anchor frame, the output forecast frame is compared with the real frame, the difference is calculated and then updated in reverse, and the parameters are iterated constantly to obtain the most appropriate anchor frame value.

Backbone is the backbone structure of YOLOv5s and consists of modules such as Focus, Conv, C3 and Spatial Pyramid Pooling (SPP). The Focus module will input data segmentation is 4, then joining together, and then for convolution operation. Focus module reduces the cost of convolution, and its main function is to reduce the amount of floating point computation and improve the running speed of the model. Specific operations are shown in Fig.5. Conv is the basic convolution unit of YOLOv5s, which is composed of conv layer, BN layer and SiLu activation function. Bottleneck refers to residual components. SPP is a space pyramid pool module. SPP performs three maximum pooling operations of different sizes on the input, and Concat concatenates the output results. The output depth is the same as the input depth.

https://doi.org/10.58291/ijec.v2i1.80

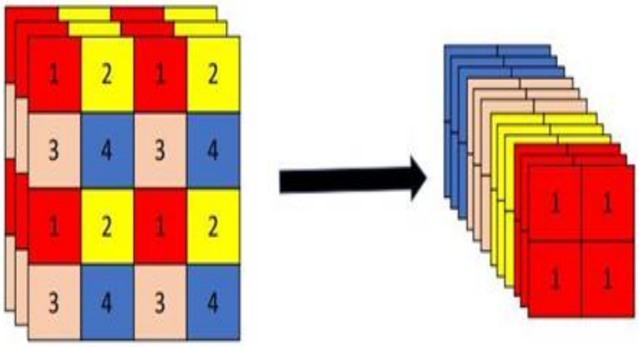


Figure 4 Focus slicing operation

In Neck, a structure combining feature pyramid and path aggregation network is adopted. The conventional top-down feature pyramid layer is combined with bottom-up path aggregation network, and the extracted semantic features and location features are integrated, which can help the model learn features better and enhance the sensitivity of the model to small targets. The Head outputs a vector with the category probability of the target object, the object score, and the position of the bounding box for that object.

Result and Discussion

At present, there are mainly two methods for lightweight network design. One method is to do some cutting and quantification on the basis of trained network, such as model pruning, weight quantification and knowledge distillation. Another approach is to design efficient neural network architectures such as MobileNetv1-v3, ShuffleNetv1-v2 and GhostNet. The design process of efficient neural network architecture is mainly to reduce convolution parameters. In the design process, on the one hand, small convolution kernel is used to replace large convolution kernel, or large convolution kernel is divided into small convolution kernel of different sizes in stages. On the other hand, by reducing the number of channels in the feature graph, the number of parameters and computation amount of the model are reduced.

MobilNet network structure

The core idea of MobileNet is that Depthwise convolution is used as a method instead of Standard convolution. Compared with the standard convolution, the depth-separable convolution can greatly compress the number of model parameters, while maintaining the accuracy less different from that of the standard convolution. Deep separable convolution, as shown in Fig.5, is decomposable convolution operation that can be decomposed into two smaller operations: deep convolution and point-by-point convolution.

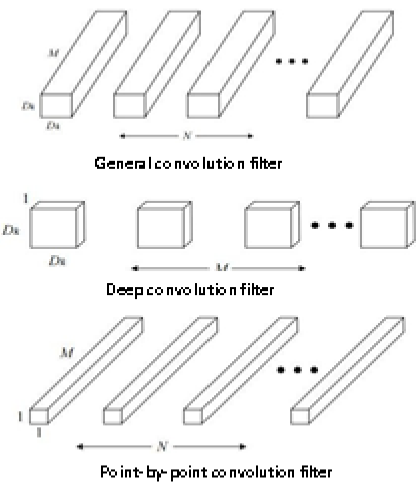


Figure 5 Common convolution filter, deep convolution filter and point by point convolution filter

The biggest feature of MobileNetv1 network is the use of deep separable convolution to reduce the number of parameters and the amount of computation, but the network structure does not adopt the residual connection. In 2018, Google launched MobileNetv2, based on MobileNetv1. MobileNetv2 is innovative enough to Inverted Residuals to run the network, as shown in Fig.6.

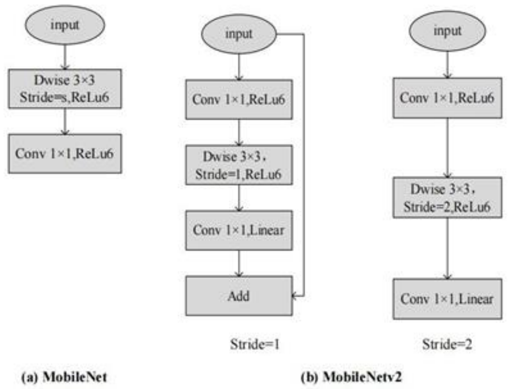


Figure 6 Mobile Net and Mobile Netv2 network and block diagram

In 2019, Google proposed MobileNetv3 network architecture, which is more accurate and real- time than MobileNetv2. The network structure continues the deep separable convolution idea of MobileNetv1 and the inverted residual structure of MobileNetv2. Unlike MobileNetv2, MobileNetv3 adds a 5×5 depthizable convolution structure and SENet (Squeeze and Excitation Networks) structure. Excitation Networks h-swish activation function is used in the nonlinear activation part.

Conclusions

The one-stage target detection architecture based on location regression directly carries out classification and location regression, which improves the speed of target detection, and the performance of missed detection rate and detection accuracy is also continuously improved in subsequent versions. Generally speaking, in order to facilitate the deployment of the algorithm on mobile devices, there are two main ways to improve the target detection algorithm in the first stage: lightweight network design and precision improvement design. Lightweight network is mainly achieved by reducing the party parameters in the model, and precision improvement can be achieved by introducing attention module. Even so, there is still a lot of work to be done in the field of detection that needs to be innovated or improved.

1. The object detection method proposed in this paper only adopts the strategy of changing the network model to reduce the computational cost. In the future, we can try to complete the lightweight design of the network model by means of model pruning and other compression models.
2. Weakly supervised learning: Large-scale instance data annotation is a very expensive and time-consuming project. By combining weak supervision with large-scale image-level classification and small-scale instance data annotation to train network models, the tagging cost can be greatly reduced
3. 3D target detection: In automatic driving and other fields, 2D images do not contain depth information, which makes it impossible to effectively avoid collision. Although there have been a lot of studies on 3D target detection, there are still many problems in practical application.

References

Ostanin, M., Zaitsev, S., Sabirova, A., & Klimchik, A. (2022). Interactive Industrial Robot Programming based on Mixed Reality and Full Hand Tracking. *IFAC-PapersOnLine*, *55*(10), 2791-2796. https://doi.org/https://doi.org/10.1016/j.ifacol.2022.10.153

Petrović, M., Jokić, A., Miljković, Z., & Kulesza, Z. (2022). Multi-objective scheduling of a single mobile robot based on the grey wolf optimization algorithm. *Applied Soft Computing*, *131*, 109784. https://doi.org/https://doi.org/10.1016/j.asoc.2022.109784

Raikwar, S., Fehrmann, J., & Herlitzius, T. (2022). Navigation and control development for a four-wheel-steered mobile orchard robot using model-based design. *Computers and Electronics in Agriculture*, *202*, 107410. https://doi.org/https://doi.org/10.1016/j.compag.2022.107410

Raudmäe, R., Schumann, S., Vunder, V., Oidekivi, M., Nigol, M. K., Valner, R., Masnavi, H., Kumar Singh, A., Aabloo, A., & Kruusamäe, K. (2023). ROBOTONT – open-source and ROS-supported omnidirectional mobile robot for education and research. *HardwareX*, e00436. https://doi.org/https://doi.org/10.1016/j.ohx.2023.e00436

Shi, K., Wu, Z., Jiang, B., & Karimi, H. R. (2023). Dynamic path planning of mobile robot based on improved simulated annealing algorithm. *Journal of the Franklin Institute*, *360*(6), 4378-4398. https://doi.org/https://doi.org/10.1016/j.jfranklin.2023.01.033

Sun, Y., Wang, X., Lin, Q., Shan, J., Jia, S., & Ye, W. (2023). A high-accuracy positioning method for mobile robotic grasping with monocular vision and long-distance deviation. *Measurement*, *215*, 112829. https://doi.org/https://doi.org/10.1016/j.measurement.2023.112829

Suwoyo, H., Abdurohman, A., Li, Y., Adriansyah, A., Tian, Y., & Ibnu Hajar, M. H. (2022). The Role of Block Particles Swarm Optimization to Enhance The PID-WFR Algorithm. *International Journal of Engineering Continuity*, *1*(1), 9-23. https://doi.org/10.58291/ijec.v1i1.37

Suwoyo, H., & Harris Kristanto, F. (2022). Performance of a Wall-Following Robot Controlled by a PID-BA using Bat Algorithm Approach. *International Journal of Engineering Continuity*, *1*(1), 56-71. https://doi.org/10.58291/ijec.v1i1.39

Suwoyo, H., Hidayat, T., & Jia-nan, F. (2022). A Transformable Wheel-Legged Mobile Robot. *International Journal of Engineering Continuity*, *2*(1), 27-39. https://doi.org/10.58291/ijec.v2i1.80

Suwoyo, H., Thong, Z., Tian, Y., Adriansyah, A., & Hajar, M. H. I. (2022). THE ACA-BASED PID CONTROLLER FOR ENHANCING A WHEELED-MOBILE ROBOT. *TEKNOKOM*, *5*(1), 103-112.

Tan, S., Yang, J., & Ding, H. (2023). A prediction and compensation method of robot tracking error considering pose-dependent load decomposition. *Robotics and Computer-Integrated Manufacturing*, *80*, 102476. https://doi.org/https://doi.org/10.1016/j.rcim.2022.102476

Yang, Y., Qin, S., & Liao, S. (2023). Ultra-chaos of a mobile robot: A higher disorder than normal-chaos. *Chaos, Solitons & Fractals*, *167*, 113037. https://doi.org/https://doi.org/10.1016/j.chaos.2022.113037

Yu, Z., Yuan, J., Li, Y., Yuan, C., & Deng, S. (2023). A path planning algorithm for mobile robot based on water flow potential field method and beetle antennae search algorithm. *Computers and Electrical Engineering*, *109*, 108730. https://doi.org/https://doi.org/10.1016/j.compeleceng.2023.108730

Zhang, D., Luo, R., Yin, Y.-b., & Zou, S.-l. (2023). Multi-objective path planning for mobile robot in nuclear accident environment based on improved ant colony optimization with modified A∗. *Nuclear Engineering and Technology*, *55*(5), 1838-1854. https://doi.org/https://doi.org/10.1016/j.net.2023.02.005

Zhang, Z., Xiao, J., Liu, H., & Huang, T. (2022). Base placement optimization of a mobile hybrid machining robot by stiffness analysis considering reachability and nonsingularity constraints. *Chinese Journal of Aeronautics*. https://doi.org/https://doi.org/10.1016/j.cja.2022.12.014

Zhao, X., & Cheah, C. C. (2023). BIM-based indoor mobile robot initialization for construction automation using object detection. *Automation in Construction*, *146*, 104647. https://doi.org/https://doi.org/10.1016/j.autcon.2022.104647