Target part detection based on improved SSD algorithm

Yongwei Yu, Xin Han, Liuqing Du*

College of Mechanical Engineering, Chongqing University of Technology, Chongqing, 400054, China

*Corresponding author’s e-mail: lqdu@126.com

Abstract: Aiming at the problem that the traditional target detection algorithm cannot balance the target detection accuracy and real-time detection, and the detection effect is poor under the actual complex production conditions. A deep learning detection method based on Inception-SSD framework is proposed. The Inception-SSD, which integrates the Inception prediction structure, introduces the Inception block into the extra layer of the SSD network. Then it uses batch normalization (BN) and residual structure connection to capture more target information without increasing the complexity of the network. It will improve its accuracy without affecting its real-time detection. The result of the experiment show that the detection accuracy of this model is 97.8% in the actual production process, which is largely improved comparing with the original SSD algorithm. The detection time is 48ms, which improves the detection accuracy and ensures the real-time performance. It can meet the detection requirements of parts in actual production.

1. Introduction

Part detection and location is an important part of the vision task of flexible assembly robots. Traditional image detection methods mainly adopt artificial design features, such as SIFT features [1], LBP [2] features, HOG [3] features. The method of artificial design feature lacks robustness to feature diversity change, and which cannot meet the requirements of modern industrial detection.

In recent years, deep learning methods have shown superior performance in natural image detection, voice processing and language recognition. Deep convolutional neural networks [4-7] can improve the classifier model and train the target features autonomously, therefore, the generalization of the trained features is much higher than that of the artificial features. At present, there are two main target detection algorithms based on convolutional neural networks. One is region-based target detection algorithms, such as Faster-RCNN [8], Mask-RCNN [9]. This kind of algorithm has high detection accuracy, but slow detection speed, and which cannot achieve real-time detection. The other is based on regression target detection algorithms, such as YOLO [10], SSD [11]. This kind of algorithm has fast detection speed and can meet real-time detection requirements, but the detection accuracy is slightly inferior. In order to overcome the defect of traditional methods, a part detection method based on Inception-SSD framework is proposed. Firstly, the Inception block is introduced into the extra layer of SSD. Then the Inception-SSD framework is established to increase the width of neural network and the adaptability to different scale features, which realizes the fusion of different scale features, and thus largely improves the recognition accuracy of the model. After that the loss function and non-maximum suppression are optimized to improve its detection accuracy in complex environment without affecting its detection speed.
2. SSD algorithm principle

SSD (Single Shot Multibox Detector) algorithm takes VGG16 as the basic network, and its model structure is shown in figure 1. The size of the input image needs to be transformed into 300 × 300, and then input it into the SSD network. SSD is based on the VGG-16 convolutional neural network, and the fully connection layers (fc6 and fc7) in VGG-16 are changed into convolutional layers (conv_fc6 and conv_fc7). Then convolutional layers of different sizes are added later to predict multi-scale targets, as shown in figure 1. Conv_fc7, conv8_2, conv9_2, conv10_2 and conv11_2 are added to the conv4_3 of VGG-16 network, and SSD networks have different levels of feature maps, which are respectively used for border deviation and category prediction of different scale targets. Finally, the detection results are obtained by non-maximum suppression (NMS).

![SSD network structure](image)

**Figure 1. SSD network structure**

SSD adopts the method of full convolution direct regression prediction, and no regional candidate box is generated, which greatly improves the detection speed of SSD. However, in the face of some complex situations such as visual interference, small parts with no obvious surface features, parts stacking, and similar appearance of different types of parts in the industry. The SSD algorithm sometimes has missed and false detection, which is inevitable in our actual factory production environment.

3. Part detection method based on deep learning model Inception-SSD

The extra layer of SSD algorithm contains limited small target information, and it only has 3 × 3 convolution kernel, then there is a problem of missing target details. At the same time, the SSD algorithm network has a complex neural network model structure. To improve its network performance, the general method is to increase the network width and depth. However, it will lead to an increase in the number of network parameters, and resulting in an increase in the amount of computation and the problem of over-fitting. Therefore, in this paper, the Inception-SSD network structure is proposed. Which is introduced into the SSD network, and the detection accuracy of the network is greatly improved.

The Inception architecture has the high performance of dense matrix and keeps the sparse network structure, which can solve the problem of over-fitting and increased computation during the optimization of neural networks. GoogLeNet Inception V1 [12] proves that to solve the problem of increased computation and over-fitting caused by the increase of parameters, the basic method is to convert the full connection layers or even half of the convolution layers to sparse links. Thus, the Inception network can capture more target information without increasing network complexity.

In order to retain the original details of the target, the Inception block is inserted to replace some extra layers in the original SSD network, and the Inception block is optimized to decompose the 5x5 convolution kernel into two one-dimensional convolutions (1x5 and 5x1). Thus splitting 1 conv into 2 conv, it further deepens the network and increases the nonlinearity of the target detection network. The redundant computing power is used to deepen the network, at the same time, the number of feature
maps in each layer of the perception network is reduced, so that the sum of feature maps is the same as that of the original SSD network. In order to reflect the significance of various scales of the receptive fields, the three convolution kernels (1×1, 3×3, 5×5) outputs are weighted separately, and the weighted value w = {1, 2, 1}. In the Inception module, each volume convolution layer is followed by Batch Normalization (BN). In addition, we add Conv1×1 convolution layer near the output to reduce the number of parameters and improve the calculation speed. Figure 2 shows the optimized Inception block.

![Inception block with BN and residual structure](image)

Figure 2. The inception building block with BN and residual structure

We replace the Conv6, Conv7 and Conv8 layers of the original SSD structure with Inception block. Specifically, each of them is replaced by three convolution towers. Conv4_3, FC7, Conv6_2, Conv7_2, Conv8_2, and Conv9_2 are used as feature extraction layers to detect target. However, as the network goes deeper, it becomes more difficult to get converged. In order to solve this problem, we introduce the residual structure in the extra layer. We concatenate the inputs and outputs of Conv6_2 as the inputs to Conv7_1.

Model based on Inception-SSD structure is shown in figure 3. Inception-SSD adopts the full convolution method, although the receptive field is the whole image, it is still need to classify each receptive field for different scales and shapes. Inception-SSD will directly predict its category and location information by regression, and it is realized by scattering in different layers. Thus it improves the network overall accuracy and speed, make full use of the advantage of dense matrix structure. Then the detection accuracy of small targets is improved when the neural network is optimized.

![The architecture of Inception-SSD network](image)

Figure 3. The architecture of Inception-SSD network
4. Experiment and analysis
In the test of part detection effect, This paper specially simulates the complex situation in the actual production, such as many kinds of parts stacking and blocking, parts assembly, unrelated debris interference, changing camera angle. The detection effect is shown in figure 4, the detection accuracy of this paper is over 97.5%, and it can meet the parts detection requirements of actual production environment.

![Detection Effect Image]

Figure 4. The detection effect of multi-class parts in hybrid stacking and assembling state

5. Conclusion
This paper proposes a method of parts detection based on Inception-SSD framework. In the original SSD network extra layer, we add the Inception and batch normalization (BN) block, and connect them with the residual structure to improve the detection accuracy of SSD network without affecting its detection speed. At the same time, we add the exclusion loss into the loss function to improve the detection effect of the model in the case of stack and occlusion. A non-maximum suppression method based on weighting algorithm is proposed to improve SSD network's ability of feature delivering. In the training stage, the best part detection model is selected by combining the loss value curve and different iteration with recall rate testing. The result of the experiment show that, under the condition of non-structural complexity, the method in this paper has significantly improved the effect of detection accuracy comparing with the original SSD network, and it can meet the requirements of actual production environment for parts detection.

Acknowledgments
Thanks to the funding of Chongqing Basic and Frontier Research Project (CSTC2017JCYJAX0344).

References
[1] Yuan, J., Cui, K., & Li, B. (2015). Identification and location of insulator video images based on ASIFT algorithm. Electrical Measure and Instrumentation, 7(7), 106-112.
[2] Yu, S., & Guiquan, L. (2015). Face Recognition Method Based on HOG and LBP Feature. Computer Engineering, 41(9), 205-208.
[3] Wang, D., & Zhang, X. H. (2013). Biomimetic recognition method of human behavior based on HOG and SVM. Journal of Jilin University: Engineering and Technology Edition, 43(S 1), 489-492.
[4] Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.
[5] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.
[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
[8] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object
detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[9] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).