OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Mechanical Part Recognition Based on Improved Faster R-CNN (Region Proposals with Convolutional Neural Networks)

Wenxuan Wu, Yixin Wang, Yuchen Bai
Southwest Jiaotong University, Chengdu 610031, China

Abstract: *With the continuous expansion of the production scale and the increasing demands for production efficiency and quality, the traditional method of manually sorting parts faces many challenges, such as low efficiency, high error-proneness, and high labor costs. In this paper, an improved algorithm for object recognition based on Faster R-CNN is proposed. The CBAM (Convolutional Block Attention Module) attention mechanism is introduced into the ResNet50 image extraction network. It can capture the important information in the feature map more comprehensively and improve the detection performance. The results indicate that both the traditional Faster R-CNN and the improved Faster R-CNN have been successfully applied to the recognition of mechanical part images using a self-made datasets of mechanical part images. Compared with the traditional Faster R-CNN, the prediction precision of mechanical parts recognized by the improved Faster R-CNN is increased by 3.7%.*

Keywords: *Faster R-CNN; ResNet50; object recognition*

Date of Submission: 16-03-2025 Date of acceptance: 01-04-2025

I. Introduction

In the field of machinery, whether it is the assembly of robot parts or the classification of parts in waste recycling stations, there are problems such as a wide variety and large quantity of mechanical parts, a single traditional classification environment, and low classification precision. Object recognition is one of the important applications of computer vision in the field of machinery. With the development of the foundation of mechanical processing automation, workers are gradually liberated from heavy labor, improving work efficiency and reducing production costs [1].

In reference [2], a fast part recognition method based on multiple contours was proposed, which compares and recognizes the images of the tested part with the template part images in the part library from three aspects: the number of contours of the part, the geometric and shape features of each contour. In reference [3], a fast method for parts was proposed, which extracts geometric information from mechanical part images using image processing techniques and interprets them as features to describe the parts. This feature description was stored as an encoding matrix, and this method was applicable to parts with simple features such as holes, steps, taper, etc.

The concept of deep learning originated from the study of artificial neural networks, which is a method of combining extracted simple features to form abstract high-level features. Since the convolutional neural network framework was proposed by Professor Hinton's team in 2012, deep learning has continued to develop and is widely used in various fields of life such as image retrieval, speech recognition, and intelligent robots [4-8].

In order to improve the precision of mechanical part object detection, a mechanical part detection method based on improved Faster R-CNN is proposed. This method incorporates the CBAM attention mechanism into the image extraction network structure of the Faster R-CNN to capture detailed features of mechanical parts and improve prediction precision.

II. Related Work

2.1 Faster R-CNN

Since convolutional neural networks (CNN) won the championship at the ImageNet Large Scale Visual Recognition Challenge in 2012, Girshick et al. first proposed the region proposals with CNN (R-CNN) framework in 2014. They used the selective search (SS) algorithm to extract candidate regions, used CNN to extract region features for classification, and performed bounding box regression based on error analysis [8-10]. However, even though selective search algorithms are used in R-CNN to extract candidate regions, there are still a lot of repetitive calculations when extracting features from all candidate regions, which affects the training speed. To address this issue, Girshick et al. borrowed the idea of spatial pyramid pooling (SPP) network in 2015 and introduced region of interest (ROI) pooling layers into R-CNN. They proposed Fast R-CNN, which extracts features from the entire image, selects regions of interest as candidate regions on the feature map, and uniformly inputs the size of the candidate regions to the next layer [11-12]. Fast R-CNN had made improvements in feature extraction, accelerating the training speed of the network. However, when extracting candidate regions, it still used a selective search algorithm, which generates over 2000 candidate regions. This was still a very time-consuming task. In 2016, on the basis of Fast R-CNN, Ren et al. abandoned the previous selective search algorithm and proposed a combination of Fast R-CNN and RPN called Faster R-CNN, which is a faster region proposal network. They introduced region proposal network (RPN) to directly generate target candidate regions [13].

2.2 ResNet50 Network structure

ResNet50 is an important member of the residual network structure series and is currently one of the most widely used models in the field of image recognition. It performs particularly well in tasks such as classification, object detection, and semantic segmentation, effectively overcoming the problem of gradient vanishing. The model takes the global average pooling layer as the key processing step before output, and uses large-scale datasets such as ImageNet for pre-training to obtain initial weights. The ResNet50 structure is divided into four main parts, each referred to as a stage, each responsible for different functions (Fig .1).

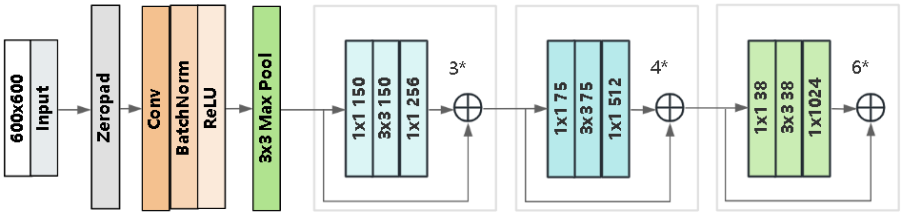

Fig.1. ResNet50 network structure

ResNet50 uses a 7×7 convolutional kernel combined with a 3×3 max pooling operation in the first stage, effectively reducing the size of the input image and providing a more compact data representation for subsequent processing. In the second stage, ResNet50 utilizes a series of innovative residual structure designs, including Conv2, Conv3, Conv4, and Conv5 residual blocks, to enable the model to explore advanced features of images in depth. Ultimately, these highly abstract features are fed into the fully connected layer of the third stage to complete the final classification task.

III. Improved ResNet50 Network Structure

3.1 CBAM Attention Mechanism

The CBAM attention mechanism is used to enhance the performance of convolutional neural networks by extracting and utilizing important feature information through channel and spatial attention modules. Its structure is shown in Fig.2. The CBAM channel attention module focuses on the importance of each channel, while the spatial attention module focuses on the importance of different positions, allowing the network to selectively enhance or suppress the feature responses of different channels and positions. The CBAM module can be embedded into common convolutional neural network structures, which can significantly improve network performance [14].

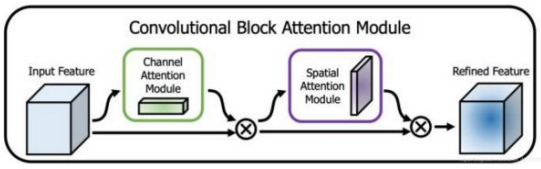

Fig.2. CBAM attention mechanism

The implementation of the channel attention module is shown in Fig.3, which can be divided into two parts. First, global average pooling and global maximum pooling are performed on a single input feature layer (input feature F), resulting in two feature strips MaxPool and AvgPool with a length equal to the number of channels in the feature layer. Afterwards, the results of average pooling and max pooling are processed using two shared fully-connected layers, the two processed results are added, and then a sigmoid is taken. At this point, the weight of each channel in the input feature layer (between 0-1) is obtained. After this weight is obtained, the channel attention Mc is obtained, and finally this weight is multiplied by the original input feature layer.

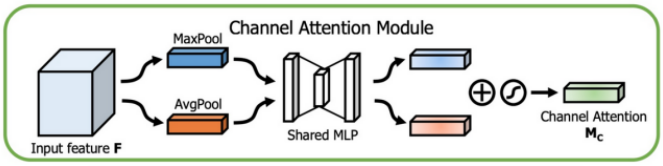

Fig.3. Channel attention module

The implementation of the spatial attention module is shown in Fig.4. First, the maximum and average values of the input feature layer Channel-refined feature F' on each feature point channel are taken. Afterwards, these two results are stacked and a convolution with 1 channel is used to adjust the number of channels. Then, a sigmoid is taken and the weight of each feature point in the input feature layer (between 0-1) is obtained. After this weight is obtained, a feature layer Spatial Attention Ms with a height×width×1 is obtained, and finally this weight is multiplied by the original input feature layer.

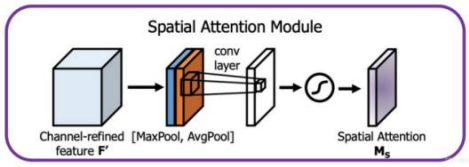

Fig.4. Spatial attention module

3.2 Improved ResNet50 Network Structure

Improvements have been made to the original ResNet50 network in this paper, and the improved ResNet50 network structure is shown in Fig.5. A 600×600pixel mechanical part image is taken as the input of the network, Batch Normalization is used to accelerate the convergence speed, and rectified linear unit (ReLU) activation function is used to increase nonlinearity. The max pooling layer is sampled through and then four ResNet50 bottleneck residual networks are sequentially passed through, each of which contains 3, 4, 6, and 3 residual structures. For each ResNet50 bottleneck module, the input feature map is first passed through a 1×1 convolutional layer to adjust the number of channels in the feature map. Perform multi-scale feature extraction and spatial transformation through a 3×3 DCNv2 layer. After passing through a 1x1 convolutional layer again, further adjust the number of channels or fuse features. Enhance through CBAM attention mechanism. Finally, all processed features are fused through an addition operation to form the final output feature map. Similarly, after passing through the second, third, and fourth residual blocks, global average pooling is performed to reduce the feature map to 1×1; Finally, a fully connected (FC) layer is used for classification, and the final classification result is output through Softmax.

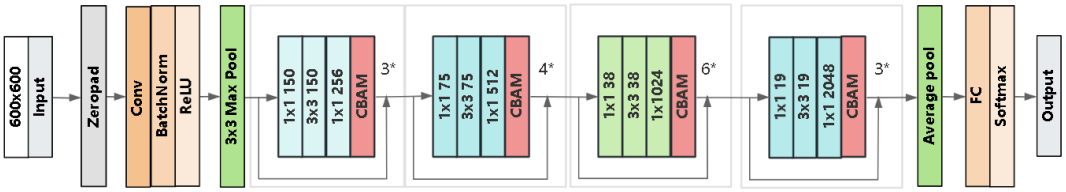

Fig.5. Improved ResNet50 network structure

IV. Experiments

4.1 Datasets

A self-made mechanical image datasets is used for training recognition. The datasets include five categories: hex screw, round head screw, locating pin, hex nut, and cap nut. The datasets consist of a total of 100 images, and the categories of mechanical parts are shown in Fig.6. LabelImg is an open-source image annotation tool that can draw frames and label corresponding categories on images. It is written in Python and Qt, and its annotation information is automatically converted into XML format, which is the same as the XML format used in the PASCAL VOC datasets and ImageNet datasets. Use LabelImg to create XML labels for the collected mechanical part images. Save the collected images and created labels in the designated folder. All images are randomly divided into training and testing sets in a ratio of 8:2.

Fig.6. Categories of mechanical parts. (a)hex screw (b)round head screw (c)locating pin (d)hex nut (e)cap nut

4.2 Experimental Environment and Evaluation Indicators

The software and hardware environment of the experiment includes:

a) Hardware environment: The CPU is Intel (R) Core (TM) i5-9300H CPU @ 2.40GHz 2.40 GHz, and the GPU is NVDIA GeForce GTX 1650.

b) Software environment: The programming language is Python 3.9 and the deep learning framework is PyTorch 2.1.0. The software environment is CUDA12.3.

The ResNet50 network in this article adopts the Adam (Adaptive Moment Estimation) optimization algorithm, with a learning rate of 0.01, weight decay coefficient of 0.0001, momentum coefficient of 0.9, model data batch size of 32, and training epochs of 400. Precision, recall, and F1-score are used as indicators to evaluate the performance of the model. The predictive performance of each category in the datasets is taken into account.

Precision: The proportion of true cases to all true cases is defined as follows:

$$V_P = \frac{V_{TP}}{V_{TP} + V_{FP}} \tag{1}$$

Recall: The ratio of the number of correctly classified samples to the total number of samples in all data is defined as follows:

$$V_R = \frac{V_{TP}}{V_{TP} + V_{FN}} \tag{2}$$

F1-score: Taking into account the precision and recall rate of the model, the formula is defined as follows:

$$V_{F1-score} = (1+\beta^2)\frac{(V_P \times V_R)}{\beta^2 \times V_P + V_R} \quad (3)$$

Among them, V_{TP} (True Positive) represents the number of samples that are truly positive categories and predicted as positive categories by the model;
V_{FP} (False Positive) represents the number of samples that are actually negative categories but incorrectly predicted as positive categories by the model;
V_{FN} (False Negatives) represents the number of samples that are actually positive categories but incorrectly predicted as negative categories by the model;
The value of β is 1.

4.3 Experiments on Faster R-CNN

After being trained and tested with the traditional Faster R-CNN, recognition results for part images were obtained, with some parameters shown in Table 1. After the average of the parameters calculated for each part in the table is taken, an average recognition precision of 93.2%, an average recall rate of 81.7%, and an average F1-score of 0.871 based on Faster R-CNN can be obtained.

Table 1
Test results of Faster R-CNN

category	Precision	Recall	F1-score
hex screw	0.952	0.821	0.882
round head screw	0.897	0.815	0.854
locating pin	0.948	0.819	0.878
hex nut	0.929	0.817	0.869
cap nut	0.935	0.816	0.871
Sum	0.932	0.817	0.871

4.4 Experiments on Improved Faster R-CNN

The improved Faster R-CNN network structure was trained using the same self-made datasets, and the change in loss values during the training process is shown in Fig.7.

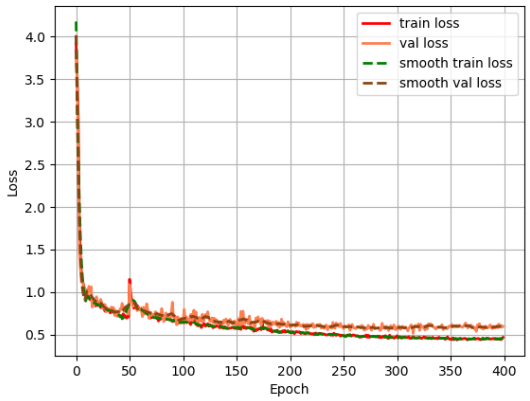

Fig.7. The process of loss value changes

The improved Faster R-CNN network structure was tested using the same images, and the test results are shown in Fig.8 and Fig.9.

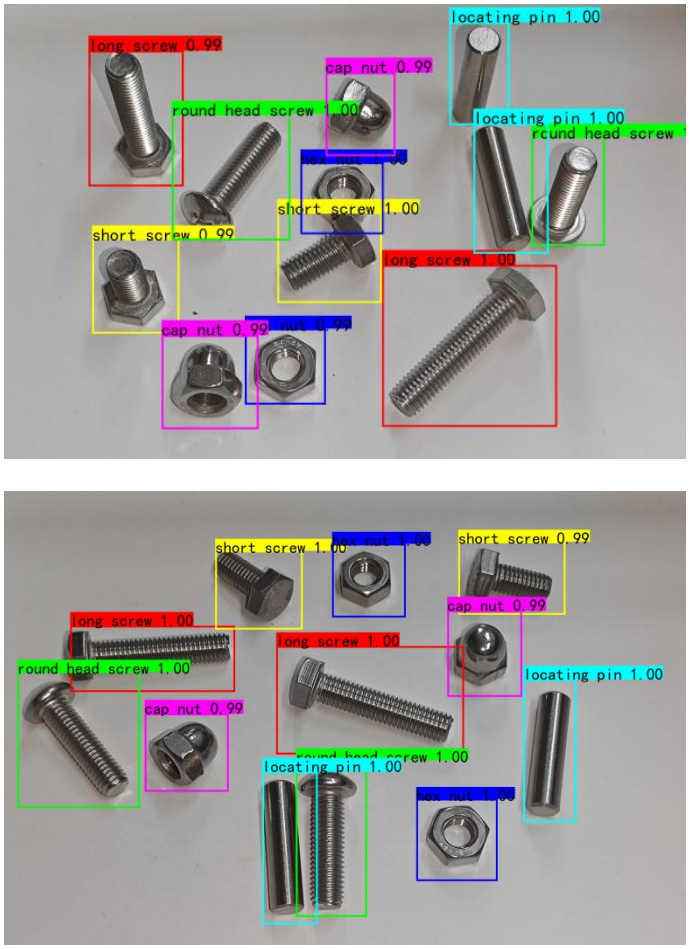

Fig.8.9. Test results

Meanwhile, the partial recognition parameters obtained by them are shown in Table 2. Taking the average of the parameters calculated for each part in the table, an average recognition precision of 96.9%, an average recall rate of 83.9%, and an average F1-score of 0.899 for part images based on the improved Faster R-CNN can be obtained.

Table 2
Test results of improved Faster R-CNN

category	Precision	Recall	F1-score
hex screw	0.983	0.842	0.907
round head screw	0.951	0.837	0.890
locating pin	0.979	0.841	0.904
hex nut	0.969	0.835	0.897
cap nut	0.964	0.839	0.897
Sum	0.969	0.839	0.899

4.5 Experiments Results and Analysis

To verify the effectiveness of the improved network proposed in this paper, the improved network and the original network were trained on self-made image datasets, and all trained networks were tested using the same test set. The comparison of the results of the two experiments is shown in Table 3. Compared to the Faster R-CNN without introducing CBAM attention mechanism, the Faster R-CNN with CBAM attention mechanism performs better in precision, recall, and F1-score. Moreover, the improved Faster R-CNN has increased recognition precision by 3.7%.

Table 3
Comparison of results between two experiments

category	Faster R-CNN	Improved Faster R-CNN	increment
Precision	0.932	0.969	0.037
Recall	0.817	0.839	0.022
F1-score	0.871	0.899	0.028

When testing the final training framework, it was found that the model trained on a datasets of mechanical part images with simple backgrounds could only recognize mechanical part images with the same simple background. When being tested with complex backgrounds, it was found that images in complex backgrounds could not be recognized by it. Only by training with images with complex backgrounds can the obtained model recognize mechanical part images in complex environments.

V. Conclusion

This paper proposes a mechanical part image attribute prediction method that integrates attention mechanism and improved Faster R-CNN. This method introduces the CBAM attention mechanism, enhances the model's representational ability, and thus improves the model's predictive performance.

The main conclusion is that with the introduction of attention mechanism, the improved Faster R-CNN model has a 3.7% higher recognition precision in mechanical part images than the Faster R-CNN model without attention mechanism.

In future research, higher quality datasets can be built for training, or separate studies can be chosen to be conducted in specific complex environments to improve the predictive precision of the model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1]. Yan K, Su Z, Huang M. Application of Support Vector Machine in Mechanical Part Recognition [J]. Application of Electronic Technique, 2008, 34(11): 108 - 110.
[2]. Si X, Wu W, Sun Y. Part Recognition and Location Based on Vision [J]. Modular Machine Tool & Automatic Manufacturing Technique, 2016(10): 70 - 73.
[3]. Chawla P R, Deb S. A Computer-Aided Inspection Methodology for Mechanical Parts Based on Machine Vision[C]//Proceedings of 4th Int. and 25th AIMTDR Conference, 2012:40-49.
[4]. Law H, Deng J. CornerNet: Detecting Objects as Paired Keypoints[C]//Proceedings of the European Conference on Computer Vision, 2018:734-750.
[5]. Wang H, Wang Y, Zhou Z, et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:5265-5274.
[6]. Johnson J, Alahi A, Fei-Fei L. Perceptual Losses for Real-time Style Transfer and Super-resolution[C]//European Conference on Computer Vision, 2016:694-711.
[7]. Liu G, Reda F A, Shih K J, et al. Image Inpainting for Irregular Holes Using Partial Convolutions[C]//European Conference on Computer Vision, 2018:85-100.
[8]. LeCun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989,1(4): 541-551.
[9]. Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.
[10]. Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.
[11]. He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,37(9):1904-1916.
[12]. Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015:1440-1448.
[13]. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks[C]//Advances in Neural Information Processing Systems, 2015:91-99.
[14]. Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C]// European Conference on Computer Vision. Cham: Springer, 2018: 3-19.