Detection of Metal Surface Defects Based on YOLOv4 Algorithm

Haili Zhao¹,²,a, Zefeng Yang¹,²,b*, and Jia Li¹,², c
¹Changchun University of Science and Technology
²School of Electronics and Information Engineering
aemail: 2033117168@qq.com,
bemail: 2019100538@mails.cust.edu.cn,
cemail: 2502311051@qq.com

Abstract: To solve the problem of low recognition accuracy and low defect location accuracy in traditional detection of surface defects of metal materials, this paper innovates on the basis of YOLOv4 architecture, and studies the influence of adding feature pyramid network module to different position of model neck on detection algorithm. Experiments have shown that adding the feature pyramid network (FPN) module after sampling on the neck network can enhance the feature information expression ability of the feature map originally input to the detection head in the size of 80×80 and 40×40, and achieve better detection results, and achieve better detection results. The experimental results show that adding feature pyramid network module to the neck can effectively improve the detection accuracy of the algorithm. Finally, compared with the traditional YOLOv4 network, the average recognition accuracy of this model can reach 92.5% and the recognition accuracy is improved.

1. Introduction

Metal parts will produce cracks, scratches, plaques, oxidation and other manufacturing defects in the production process, which will not only affect the appearance of the product, but also have a greater impact on the quality and service life of the product. Modern industry requires more and more high quality of metal materials, which makes the detection of metal surface defects become a research hotspot in related industries. If the defects of the metal surface can be found in time, the existing safety risks can be eliminated. The traditional manual detection method has low efficiency, high false detection rate and missed detection rate, which can not meet the production demand of modern fast-paced production line. Therefore, it is of great significance to study an efficient metal surface defect detection method.

In recent years, the deep learning model represented by convolution neural network has been successfully applied in many computer vision fields[11], the object detection method based on deep learning is widely used in many scenarios. The main depth models are one-stage network YOLO series and two-stage network R-CNN series. There are three main problems to be solved in the detection of metal defects: the environmental light is very complex, the picture has more noise, the small crack target detection problem, the cause of the defect is different, and the shape is complex and changeable. Because the industrial equipment environment is complex, the light and dark change is big, the crack is the small target, the crack may be curved, the shape is complex. These problems bring great challenges to detection, which makes most of the detection algorithms have low recognition rate and high missed detection rate. Accordingly, in order to improve the detection performance of many defects on metal...
surface, the improved YOLOv4 algorithm is used to detect defects on metal surface.

2. YOLOv4 algorithm

2.1. Algorithm Principle

Convolutional neural network is a kind of feedforward neural network with convolution operation and deep structure, which imitates the visual reception and visual cognitive mechanism of organisms[2]. YOLOv4[3] which is proposed by the Russian Alexey and is improved on the basis of YOLOv3, it belongs to a one-stage detection algorithm and is a "end" to "end" convolutional neural network defect detection model. The original intention of the YOLOv4 is to speed up the running speed of the model and optimize the neural network in parallel computing, so that the model can be trained and detected in conventional GPU.

The YOLOv4 structure diagram is shown in figure 1 and can be roughly divided into four parts: input layer, BcakBone network, Neck, Head network. The CBM in the diagram is the most integral part of YOLOv4 network structure, consists of convolution layer (Conv), batch normalization (BN), and mish activation function. The function of the convolutional layer is to extract feature information from the input image. The parameters of the convolutional layer include preset convolution kernel size and convolution operation step size and convolution kernel parameters that need to be learned by the network. For the same picture, different convolution kernels can extract different features. Therefore, in practical applications, a convolution layer usually contains multiple convolution kernels, and each convolution kernel corresponds to a feature map. Batch Normalization (BN) is a process of normalizing data, which is generally applied after the convolutional layer or the fully connected layer and before the activation function. CBL consists of convolution layer (Conv), batch normalization (BN), Leakyrelu activation function; CSPX learn from the CSPNet network structure, consisting of CBM and x residual modules. YOLOv4 adjust the input image size to 640×640 input to the network for training and detection; Its backbone network uses CSPDarknet53[4], the Darknet53 contains five large residual blocks, the number of small residual units contained in these five large residual blocks is 1, 2, 8, 8, 4, respectively. CSPDarknet53, the activation function in the first convolutional layer in the original network structure is changed from Relu to Mish, more smooth than the Relu activation function. There is no non-conductive point in the global, which has better feature transfer ability and better generalization ability. Expressions for Mish activation functions such as formula (1).

\[
\text{Mish} = x \times \tanh(\ln(1 + e^x))
\]  

(1)

The spatial pyramid pooling (SPP) block of deep convolution network is added to the CSPDarknet53. This module mainly adopts the pool operation of four maximum pool cores: 1×1, 5×5, 9×9, 13×13. In the convolution stage, the number of channels of feature maps usually rises. This will lead to a larger number of parameters. In addition, the receptive field of the high-level network layer has strong semantic information representation ability, and the receptive field of the low-level network layer has strong representation ability of spatial detail information, and the image resolution is high. Spatial pyramid pooling is one of the important measures for multi-scale pooling of high-level features in the target detection algorithm to increase the receptive field. Therefore, it is necessary to increase the pool layer between the convolution layers, on the one hand, to reduce the parameter quantity of the feature map and reduce the calculation amount; on the other hand, to increase the receptive field of the convolution kernel, so that the model pays more attention to the global features and retains some important feature information, and provides certain displacement invariance and rotation invariance for the network. At the neck of the network, the fusion of shallow features and high-level semantic features and multi-scale receptive fields is realized by PAN+SPP model structure. Using the idea of regression plus classification, the detection head divides the input images into three different size grid maps of 80×80, 40×40, 20×20, and realizes the detection of small, medium and large targets, respectively.
2.2. Indicators
In the field of target detection, in order to accurately evaluate the effect of model detection, it is often necessary to calculate the accuracy Precision rate P and recall Recall rate R. A target detection model with better performance should maintain the Precision value at a good level while the Recall value increases. A classifier with a poor training effect may be improved by losing Precision. Mean Average Precision (mAP) is an important index to measure the efficiency of target detection, which is determined by Precision rate and Recall rate respectively. The curve with Recall as horizontal axis and Precision as vertical axis is referred to as P-R curve. The product below the curve is recorded as the precision mean value. The average value of the precision mean value of all target categories is mAP value. The larger the value, the better the effect of neural network model.

In the problem of target classification, it is assumed that the target to be classified is divided into positive examples and the background is divided into negative examples. Recall rate response model recall performance, indicating that the correct number of samples predicted as the target accounts for the number of actual targets, calculated by formula (2).

\[ \text{recall} = \frac{TP}{TP + FN} \]  

Precision rate, also known as accuracy rate, response model detection ability, represents the proportion of the predicted target sample to the predicted target, calculated by formula (3).

\[ \text{precision} = \frac{TP}{TP + FP} \]  

In the formula: \( TP \) represents the number of targets detected correctly by the model; \( FP \) represents the number of targets detected by the model; \( FN \) represents the number of targets missed by the model.

2.3. Improved YOLOv4 algorithms
YOLOv4 algorithm uses CSPDarknet53 as the backbone feature extraction network and outputs three feature layers of different sizes. The feature layer enters the PANet structure after convolution operation for feature fusion and maximum pooling into the SPP structure, these practices can extract defect features of different sizes and types to some extent. However, there is a certain similarity in the defect characteristics of the metal surface. If the original YOLOv4 algorithm is directly used for training and detection, the detection effect obtained is not ideal. The defect detection system based on convolution neural network depends largely on the accurate location of defect location. The traditional YOLOv4 feature extractor deepens with the convolution of the internal system, although more advanced semantic
information is detected more, the spatial information will be lost accordingly with the decrease of spatial resolution lead to the final inaccurate positioning of the target. Therefore, inspired by the YOLOv3[5] algorithm, based on the original architecture of the algorithm, the FPN[6] module is added after the YOLOv4 Cnocat module. On the one hand, it can deepen the network depth and increase the network capacity and complexity. On the other hand, it can obtain a larger field of experience and obtain more Global and semantically higher level of feature information, so as to more effectively extract the features of the defective target, to enhance the feature information expression ability of the feature map originally input to the detection head in the size of 80×80 and 40×40, thereby improving the accuracy of target detection. The specific algorithm modification area is shown in Figure 2 below.

3. Test results and analysis

3.1. Target detection environment
The experimental platform environment is as follows: operating system Win system, CPU is Intel (R) Xeon (R) CPU E5-2678 v3, RAM is 64G. GPU is NVIDIA GeForce GTX 1080Ti. The deep learning framework is Darknet. The training settings are as follows: input image size is 640×640. The total number of incoming images per iteration is 900, momentum is 0.937. The weight attenuation coefficient is 0.0005. The initial learning rate was 0.001, batchsize is 12. Set the training Epochs to 300 times.

3.2. Results analysis
Based on Northeastern University (NEU) surface defect database, this paper selects three defect types from this database: crazing, patches, scratches, 300 pictures each, and randomly selected 720 images as a training set, using the remaining 180 images as a test set, The ratio of training set to test set is 8:2. The three defect types are shown in figure 3.
Different types of defects are marked with different color boxes, such as scratches, and the model is marked with scratches words and confidence. After training and learning the model, the partial detection effect on the test set is shown in figure 4.

When the target model detects multiple categories, the mAP of multiple categories needs to be calculated, calculated by formula (4). In the formula: C represents the number of categories.

\[
\text{mAP} = \frac{1}{C} \sum_{i=1}^{C} AP_i
\]  

(4)

When the model training is completed and the algorithm is improved and enhanced, the mAP diagram P-R graph of the improved YOLOv4 algorithm model are obtained, as shown in figs .5 and 6, respectively.

4. Conclusions
This paper proposes a surface defect detection method based on YOLOv4 depth learning network. The improved YOLOv4 model basically covers the whole coordinate system in the P-R curve of the test set
in this paper. The average detection accuracy is 92.5%, which is improved compared with the traditional algorithm. However, the training data set of this application study has only 720 picture data, which can not meet the requirements of the YOLOv4 training set. This may lead to over-fusion and misjudgment in the recognition of similar defects, which is an important reason for the low accuracy of the surface defect detection test. In the future research, we will focus on overcoming these problems to further enhance the performance and real-time performance of the detection model.

References
[1] Tao Xian, Hou Wei, Xu De. A Review of Surface Defect Detection Methods Based on Depth Learning [J/OL].]1 Journal of Automation :1-19[2021-03-08].
[2] Jin Yinggu, Zhang Tao, Yang Yanning, Wang Yue, Liu Yuting. A Review of Product Defect Detection Methods Based on Depth Learning [J.] and Journal of Dalian University for nationalities 22(05):420-427.
[3] Bochko vskiy A,Wang C Y,Liao H Y M.YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. arXiv Preprint arXiv: 2004.10934,2020.
[4] WANG C Y,LIAO H Y,WU Y H,et al.CSPNet: a new backbone that can enhance learning capability of cnn[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops.New York: IEEE,2020: 390-391.
[5] REDMON J,FARHADI A.YOLOv3: An Incremental Improvement[R]. arXiv,2018.
[6] LIN T Y,DOLLAR P,GIRSHICK R,et al.Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 2117-2125.