Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 1, No. 2, 2022

Object Detection Method of Power Equipment Based on Mask R-CNN

Chen Wang, Chunjiang Pang

School of Control and Computer Engineering, North China Electric Power University, Hebei, 071000, China

Abstract: With the rapid development of deep learning technology and its outstanding performance in the field of image, more and more researchers begin to pay attention to the application of deep learning in the field of power industry. After analyzing the structure of Mask R-CNN and considering the particularity of infrared image data set, a new Mask R-CNN model is proposed. Channel attention mechanism is introduced to make the network learn the weight coefficient of each channel, so that the network can filter noise more effectively and extract more information related to the object. Experimental results show that the accuracy of the improved model is better than that of the original model, and the effectiveness of the improved method is verified.

Keywords: Power equipment, Object detection, Mask R-CNN.

1. Introduction

In recent years, with the proposal and development of the concept of "smart grid". image processing technology has been more and more widely used in power equipment fault diagnosis. It is the first step to realize intelligent fault detection to identify, locate and classify power equipment accurately. This task can be accomplished by extracting image features. In chronological order, it can be divided into traditional image processing technology of artificial design features and feature extraction technology based on deep learning. At present, object detection technology has two main development directions: object detection algorithm based on candidate region and object detection algorithm based on regression, also known as two-stage and one-stage object detection algorithm respectively.

The two-stage object detection algorithm has obvious advantages over the first-stage algorithm in accuracy, and among the two-stage object detection algorithm, Mask R-CNN[1] has better object detection accuracy and good instance segmentation level. In this paper, Mask R-CNN is selected as the basic model of the task to realize the recognition of different kinds of power equipment in the infrared image. In this paper, the Mask R-CNN model and its improvement during the model construction will be introduced in detail, so that it can better complete the object detection task of infrared images in the electric power scene. Firstly, the overall structure of the basic model Mask R-CNN is introduced. Secondly, the design and implementation process of object detection task is introduced, including the construction of digital backbone network and the improvement of network. Finally, the experiment proves that the improved module improves the accuracy.

2. Related Works

Most of the two-stage object detection methods are completed by the following steps: Firstly, some possible regions are screened out, which are enlarged or reduced to a fixed size. Then, these regions of the same size are divided into positive samples and negative samples according to the set rules. Feature vectors are obtained by training CNN, and then they are put into the classifier to realize classification, screening and correction of generated detection boxes. Representative research methods of two-stage object detection include R-CNN[2], SPP-net[3], Fast R-CNN[4], Faster R-CNN[5] and Mask R-CNN, etc.

The object detection model adopted in this paper is Mask R-CNN, as an extension of Faster RCNN, and its structure is mainly divided into two stages. The task of the first stage is to generate the rectangular box with a high probability of detecting objects in the image, namely the region proposal box, which is called region proposal in the original work and mainly consists of convolutional neural Network and Region Proposal Network (RPN). Feature extraction of input data and suggestion box generation; The task of the second stage is to output the rectangular box containing the detection object, the corresponding prediction confidence and the mask of the generated object, including the Region of Interest (RoI) corresponding to the suggestion box, RoI is converted into Region of Interest Align (RoIAlign) layer of fixed-size feature map, and classification branch, position prediction branch and mask generation branch for classification and regression object detection frame. Compared with Faster RCNN, the improvement of Mask R-CNN lies in the increase of Mask generation branch and the replacement of RoIAlign of RoIPool. Figure 1 shows the construction of Mask R-CNN.

2.1. RPN and RoI Align

Region Proposal Network (RPN) is a classless object detector based on sliding window realized by convolutional neural Network. It first appeared in the Faster RCNN model and greatly improved the generation speed of candidate boxes. The input of RPN is the image of any scale, and the output is a series of rectangular candidate boxes (Anchor). Characteristics of each pixel point on the figure will generate a large number of candidates for the different size and different aspect ratio box, these candidate box will cover more image area, if possible by CNN to judge the candidate box which is a object is sample box, which is not contain the object negative samples, and to determine the coordinates correction sample box.

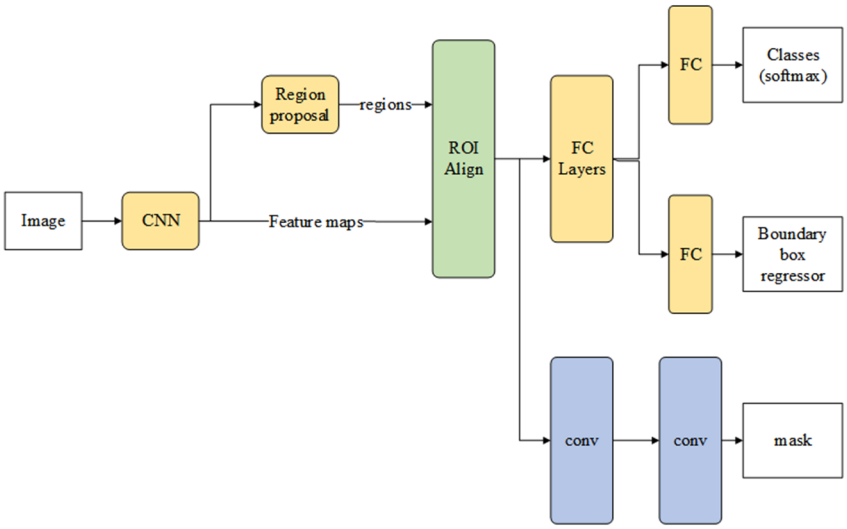

Figure 1. Mask R-CNN

In Faster R-CNN, this step corresponds to RoI Pooling layer, where the operation is: firstly, the suggestion frames are mapped back to the scale of feature maps, and then the feature map region corresponding to each suggestion frame is horizontally divided into fixed-size grids (bin), and each piece of the grid is maximized to make the corresponding regions of suggestion frames of different scales have the same resolution. In this process, both the mapping operation and the meshing operation are likely to generate floating point numbers, and the RoI Pooling layer quantifies these two steps, i.e., rounding. The RoIAlign layer is used in Mask R-CNN, and the steps of mapping and grid partitioning are no longer quantified. Floating point numbers are retained directly. Bilinear interpolation method is adopted to obtain the coordinate positions of a fixed number of sampling points in each bin, and then the maximum pooling operation is carried out to obtain more accurate features.

2.2. SENet

In recent years, attention models in different fields have been proposed, which are widely used in deep learning tasks such as natural language processing, speech recognition, object detection and image segmentation. It can be easily embedded into existing networks to improve the representation of models in a plug-and-play manner.

The attention mechanism for SENet[6] is composed of two operations squeeze and excitation. Firstly, squeeze is used to compress from the spatial dimension. Since the convolution operation is carried out in the local area, the output features cannot use the information outside the local area. This problem is more serious for the underlying network. So the squeeze operation compresses the global spatial feature on a channel into a global feature descriptor, that is, the H* W * C feature into a 1*1* C feature, using the statistics generated by the global average pooling operation.

To take advantage of the global description characteristics obtained by the squeeze operation, the nonlinear relationship between the different channels is used and the weighting is generated for each channel. In order to limit the complexity of the model and improve the generalization ability, a bottleneck structure composed of two fully connected layers was adopted. The first fully connected layer was used for dimensionality reduction and ReLU activation, and then the second fully connected layer was used to restore the original dimension. Finally, the Sigmoid function is used to obtain the normalized weight, and the final feature is obtained by multiplying the weight by the original feature.

3. Improved Mask R-CNN

3.1. Put SE into Mask R-CNN

SENet construction is very simple, does not need to introduce additional new functions or convolution layer, and has good characteristics in increasing the computational complexity. Theoretically, the additional computation amount increased by SENet is less than 1%. Therefore, the introduction of this structure into the existing network structure is very friendly to the increase of parameters and computation. Therefore, the channel attention mechanism is added to the trunk network in this task, combined with the ResNet module structure of SE.

3.2. RPN Adjustment

RPN generates candidate boxes of 9 specifications for each position on the shared feature map. The length to width ratio is 1:1, 1:2, 2:1, and the area is 128×128, 256×256, and 512×512. A total of 9 types of candidate boxes are generated by the combination of length to width ratio and area. Considering the size and proportion of power equipment objects in this task, the original aspect ratio was adjusted to 1:1, 1:3 and 3:1 to achieve a more accurate distinction between foreground and background.

4. Experiments

4.1. Experiment Platform

The following describes the construction of the experimental environment, including the development language, development environment and deep learning framework. In this experiment, Python3.7 was used as the development language, Python numpy, OpencV, Matplotlib and other third-party libraries were installed, the operating system was Ubuntu18.04, and Pytorch was used as the deep learning framework. Considering that the training process of deep learning requires high GPU memory and computing speed, NVIDIA professional computing graphics card configured with Compute Unified Device Architecture (CUDA) is used to accelerate the training of the model.

4.2. Data Set

The first step is to obtain original data, use infrared thermal imager to take infrared photos of power equipment, and screen out clearer pictures that meet requirements.

In the second step, labelme software is used to label power equipment. The photos screened in the first step mainly include circuit breakers and disconnecting switches. After labeling with labelme, a .json file corresponding to the image can be exported. The content of the file is the manually labeled point coordinates and the object category, which will be used for the training of model weight.

In the third step, use the images and .json files obtained in the second step to convert a dataset called COCO format. There are 1200 data sets, and the images are randomly divided into training set, verification set and test set according to the number ratio of 10:1:1.

4.3. Model Training Method and Parameter Setting

In the training process, the training set made in the previous section is used to train network parameters, i.e., there are 1000 images in total. Because Mask R-CNN requires massive data for training to achieve excellent performance, the number of power equipment infrared image data sets in this paper is small, so the task of this paper adopts the strategy of transfer learning to obtain the pre-training model.

An epoch refers to the process of feeding all data into a model to complete a previous calculation and back propagation. In this task, each epoch is trained on 1000 images in the training set. The initial value of the learning rate was 0.001, the batch size was 2, the step size of each epoch was 1000, and the momentum was 0.9. During the training, Adam optimizer is used to update the parameters in the model. This method updates the network parameters by randomly selecting the image gradient. Compared with the random gradient descent algorithm (SGD), it can automatically adjust the learning rate and has better performance.

In order to prevent parameter overfitting, the validation set made in the previous section was used to verify the performance of the model after the training of each epoch. Figure 2 is the curve of loss value in the training process. The experiment stopped training after iterating 150 epochs. The final test set is used to test the final accuracy of the network and output evaluation results.

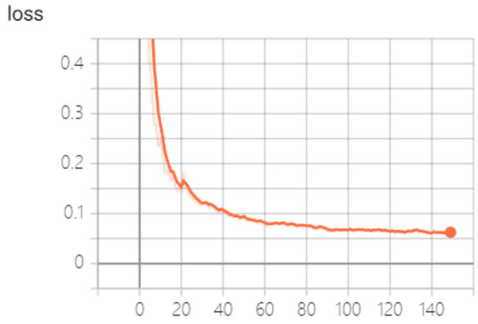

Figure 2. Loss value during training

4.4. Experimental Results and Analysis

In order to evaluate the effectiveness of the improved model in detecting objects in this task, the standard COCO index is used to measure the performance of the model. As described in the preceding section, AP50 represents an AP whose IoU threshold is set to 0.5. AP75 represents an AP whose IoU threshold is set to 0.75. AP represents the average value of aps whose IoU threshold ranges from 0.5 to 0.95.

Table 1 is the experimental results on our data set. It can be seen that the AP value of the Mask R-CNN model after the introduction of the attention mechanism is greatly improved, and the three AP values are increased by 1.7%, 2.0% and 1.8% respectively, which indicates the effectiveness of the attention module introduced in this model, with a small increase in calculation but a large improvement in performance.

Table 1. The experimental results

Methord	AP	AP50	AP75
Mask R-CNN	81.1	87.4	76.3
Improved Mask R-CNN	82.8	89.4	78.1

5. Conclusion

In this paper, the Mask R-CNN model is introduced, and an improved method is proposed for the infrared image data set of power equipment. The specific links of the experiment are introduced, including the experimental environment, the specific production process of the data set, the training method of the model and parameter setting. The effectiveness of the improved method on the data set in this paper is proved by experiments.

References

[1] HE K, GKIOXARI G, DOLLáR P, et al. Mask R-CNN; proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), F 22-29 Oct. 2017, 2017 [C].

[2] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation; proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, F 23-28 June 2014, 2014 [C].

[3] 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): 1904-16.

[4] GIRSHICK R. Fast R-CNN; proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), F 7-13 Dec. 2015, 2015 [C].

[5] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-49.

[6] LI X, WANG W, HU X, et al. Selective Kernel Networks; proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), F 15-20 June 2019, 2019 [C].