Research on Object Detection Algorithm Based on Semantic Features

Shuili Zhang[1,2,*], Xue Tian[1], Kexin Zheng[1]
[1]*College of Physics and Electronic Information, Yan'an University, Yan'An, Shaanxi, China*
[2]*Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, Yan'An, Shaanxi, China*
**Corresponding Author.*

Abstract: In order to tackle the problem of low recognition accuracy for small objects or complex backgrounds due to the single and low-resolution feature maps extracted by the traditional Faster R-CNN network, this paper designs a target detection algorithm that combines semantic features with traditional features by integrating a U-Net semantic segmentation network with the Faster R-CNN network. To create region suggestions, the extracted convolutional and semantic features are first processed using the RPN network, these proposals are then unified to the same size. After fusing the features, the candidate boxes are brought back to identify the target items. This method addresses the issue of target object detection in complicated backdrops and the loss of small target features in deep feature maps. Lastly, the network model's resilience and capacity for generalization are increased by applying the soft-NMS method to improve the detection of overlapping regions. According to experimental results, this approach performs better overall and achieves higher detection accuracy on the COCO dataset when compared to the standard Faster R-CNN network.

Keyword: Faster R-CNN; Object Detection; Semantic Segmentation; U-Net

1. Introduction

In the field of computer vision, object detection primarily focuses on identifying objects that are of interest to humans within images or videos. And at the same time detect their position and size, which should not only solve the classification problem, but also solve the positioning problem [1]. Conventional object detection algorithms primarily rely on feature extraction and classifiers to identify targets. Shayhan et al [2] proposed to use Histogram of Oriented Gradients to extract image features and adaptive background mixture models to detect pedestrians, due to its simplicity and stability, this method is still widely used, however, these traditional methods suffer from slow processing speeds and lower recognition accuracy. Through the application of deep learning, Girshick et al 2014 release of the R-CNN [3] greatly expanded the field of object detection technology; Fast R-CNN [4] was proposed In 2015, which achieved multitasking training and accelerated training and detection speed; In that same year, Ren, He and their colleagues introduced the renowned Faster R-CNN [5] algorithm, which employs a compact regional candidate network. This not only decreases the quantity of candidate boxes but also enhances detection accuracy. Wu et al [6] detect traffic signs and found that used the Faster R-CNN algorithm has good validity and robustness under the influence of different light, occlusion and motion.

For the target detection algorithm of single-stage detection, the most prominent representative is YOLO algorithm [7], which does not need to generate and classify candidate regions and has a faster detection speed. Zhang and colleagues [8] devised an apple detection model using the efficient YOLOv4 architecture, enhancing the speed of apple identification even in challenging conditions such as leafy surroundings, complex lighting, and confined spaces with high density. In 2016, Bilen et al proposed the WSDDN model [9], which uses category labels of images to train target detection tasks and greatly reduces the cost of data set preparation. However, the detection accuracy of YOLO and WSDDN algorithm is lower than that of Faster R-CNN.

At present, researchers have been solving the problem that the image information extracted from Faster R-CNN is incomplete because the target is weak and the target is blocked by a large area, resulting in low target detection accuracy. For example, Li et al [10] proposed a

MA-Res Net model, this method replaces the Faster R-CNN's original feature extractor VGG-16, which is employed for detecting small targets in 6G intelligent autonomous traffic, demonstrates superior performance in terms of stronger classification accuracy compared to other feature extraction models used for small target traffic detection. A network that FR-RCNN was proposed by Jiao et al [11] for the purpose of localizing pests on small targets, this approach produced a mAP of 56.4% on a 24-class pest dataset, which is 7.5% higher than the performance of Faster R-CNN.

In this paper, we investigate solving the problem of invisible target objects in images with very complex backdrop and the problem of feature information loss of small targets in deep feature maps using a high-level semantics of target features, therefore, a design was proposed to use Faster R-CNN to extract convolutional feature maps and a semantic segmentation network U-Net to extract semantic feature maps, and to use the RPN network to generate candidate boxes, then combine the two features using a number of dimensional transformations to determine target categorization and localization, which will enhance the target detection algorithm's capacity for generalization. It is proved through experiments that adding semantic features to the network increases feature expression ability, thereby improving detection accuracy.

2. Faster R-CNN Algorithm

Faster R-CNN's key contribution is its Region Proposal Network (RPN), which efficiently extracts candidate regions, replacing the time-consuming selective search method with a simpler approach. Figure 1 illustrates the network model structure of Faster R-CNN.

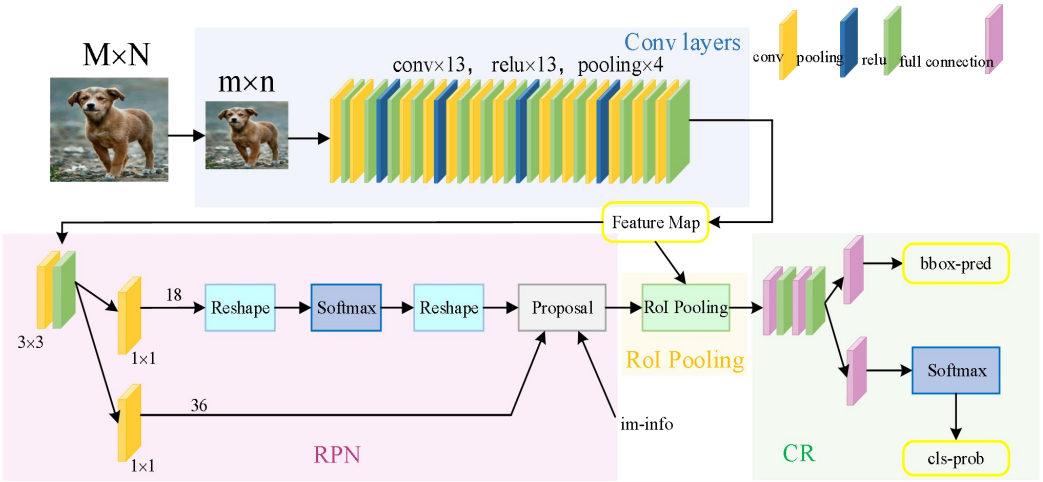

Figure 1. Structure of the Faster R-CNN Network Model

The Faster R-CNN network model is mainly divided into four parts:

(1) The input image is processed through 13 convolutional layers, 4 pooling layers, and 13 activation layers in the feature extraction layer to create the feature map. The final feature map is 1/16th of the original size, or (m/16) × (n/16).

(2) RPN (Region Proposal Network) is a region candidate network that provides a candidate frame for a classification regression network, The RPN network utilizes the feature map extracted in the initial step as input. This network traverses the feature map using a 3×3 sliding window and computes numerous anchors by aligning the center of the sliding window with the center of the original image; the upper 1×1 branch in the figure predicts whether a certain box is the background or the target through the Softmax activation function after convolution, and the lower 1×1 branch is the four offsets of the predicted box, which is regressed to the real coordinates during training to make the predicted object position more accurate in testing, and to obtain more precise candidate boxes, the data from the two branches is combined.

(3) RoI pooling is a kind of pooling layer, such as Figure 2 is the schematic diagram of RoI pooling operation. Through the RPN network, candidate boxes of the feature map are extracted, which is equivalent to matching each candidate box to different parts of the original feature map, and then clipping these sections before adding them to the ROI Pooling layer, which involves uniformly sizing the candidate boxes of various sizes.

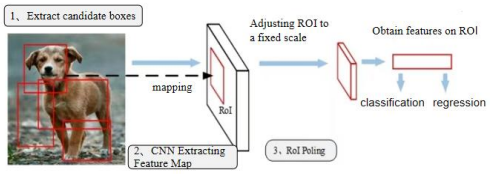

Figure 2. Schematic Diagram of RoI Pooling Operation

(4) CR (Classification and Regression), this network is used for classification and regression, the input is the candidate box area after pooling, the category and exact location of the object inside the area of interest constitute the output. To determine which exact category each candidate box belongs to, the network computes the fully connected layer and Softmax function. And outputs the probability vector and then utilizes bounded regression to get a more accurate detection box.

3. Design of a U-Net-Based Object Detection Network

Faster R-CNN mainly extracts image features through CNN. However, its feature maps are single-layered with relatively small resolutions, posing challenges in detecting background clutter and multi-scale small targets effectively. To address this, the paper integrates a semantic segmentation network with Faster R-CNN. This integration not only enhances feature information but also enhances detection accuracy.

3.1 U-Net Network

Figure 3 shows the U-Net [12] network structure. It is made up of a downsampling on the left and an upsampling on the right, following the convolutional neural network architecture, which reuses two 3×3 convolution, a linear rectification activation function (Relu) and a 2×2 max-pooling, after increasing the number of feature channels to 1024, the right upsampling is carried out, each step of up-sampling is the same as that of down-sampling, except that the max-pooling of down-sampling is changed to 2×2 deconvolution; the gray arrows depicted in the figure connect the corresponding feature maps from the left to the right side. The resulting image segmentation output represents the final segmentation result.

In this paper, in order to fuse semantic features with convolutional features in the design, the size and scale of the output feature maps of each Dconvxy module of the U-Net network are all pooled to the same size through ROI pooling as shown in Figure 4.

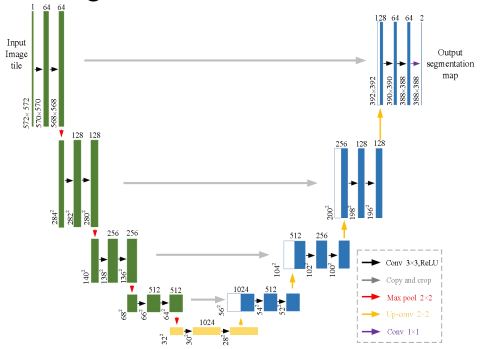

Figure 3. U-Net Network Structure

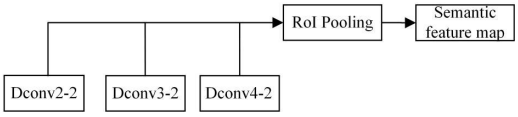

Figure 4. Multilayer Semantic Feature Extraction Example Diagram

3.2 Target Detection Algorithms Incorporating Semantic Features

The target detection algorithm designed in this paper is based on Faster R-CNN network by adding semantic segmentation network (U-Net), and its algorithm flow is shown in Figure 5.The whole algorithm includes two branches, candidate region convolution branch and candidate region semantic branch, the following is the precise algorithm flow:

(1) After inputting an image, firstly, the VGG16 network is initially employed to extract convolutional feature information. Subsequently, in order to create candidate boxes, the region candidate network module locates regions on the feature map. Following this, the candidate frames undergo ROI Pooling to standardize their sizes, resulting in transformed candidate frames. These frames are then passed to the candidate region generation module to produce the candidate region convolutional feature map, thereby forming the convolutional feature matrix.

(2) The image goes through the lower branch U-Net network to build the semantic feature map. As with the upper branch method, the candidate frame is generated by RPN module and dimensionally transformed by ROI Pooling to align it consistent with the dimensionality of the upper branch candidate area feature map, and then enter the candidate area module to generate semantic features of the candidate area and generate the semantic feature matrix.

(3) The convolutional and semantic features

provided by the upper and lower branches are fused. Finally, the fused features are fed into the candidate frame regression module for target localization and classification to output the final detection results.

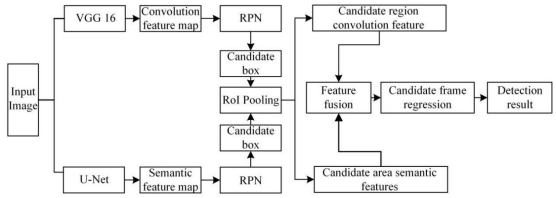

Figure 5. Flowchart of the Overall Algorithm Model

3.3 Soft-NMS Algorithm

In target detection, prediction frames produced by the model are often stacked together, and each prediction frame is assigned a score. However, the use of a sliding window may result in a significant number of prediction frames overlapping with one another, either containing or sharing most of their regions. To address this, post-processing techniques like Non-Maximum Suppression (NMS) [13] are commonly employed. The process involves sorting detection frames by their scores and retaining the highest-scoring frame. Concurrently, any overlapping boxes exceeding a certain threshold are removed, and prediction boxes meeting the criteria are selected. By removing overlap and duplication and maintaining high-scoring frames, this seeks to increase object detection efficiency and accuracy.

However, the NMS often removes non-redundant boxes in cluttered backgrounds, and its threshold is challenging to determine. Setting it too low may result in either low or high scores, while setting it too high may increase the likelihood of false detections, thus reducing detector accuracy. In order to enhance NMS, the Soft-NMS [14] approach is used. To reduce the score of adjacent boxes whose IOU is greater than the threshold value, Soft-NMS modifies the score reset function and sets a penalty function instead of immediately setting zero for all the boxes whose IOU is greater than the threshold value. Equation (1) is the scoring function of the Soft-NMS algorithm.

$$s_i = \begin{cases} s_i & \mathrm{iou}(M,b_i) < N_t \\ s_i(1-\mathrm{iou}(M,b_i)) & (1-\mathrm{iou}(M,b_i)) \geq N_t \end{cases} \quad (1)$$

Considering that the above scoring function is not continuous, a break occurs when the NMS threshold is reached. Consequently, the Gaussian weighting function is used in this paper to tackle the continuity problem, the function is represented in equation (2), where bi indicates the detection score, M indicates the box with the highest candidate score, bi is also used to refer to any other competing candidate box, IoU (*) measures either the intersection or union between two boxes, and σ indicates the variance in the Gaussian function.

$$s_i = s_i e^{-\frac{\mathrm{iou}(M,b_i)^2}{\sigma}} \quad (2)$$

4. Experimental Design

This experiment utilizes the programming language of Python and runs on a server in the lab. The hardware configuration used is a Tesla T4 GPU with 16GB of video memory; computing is based on CUDA10.2 on the operating system Ubuntu 18.04, and the deep learning framework is Tensorflow.

4.1 Datasets

Image detection may be accomplished with the COCO dataset. It contains more than 330,000 photos with 1.5 million targets spread throughout 80 target categories, there are 91 material classifications, and each image contains five sentences describing the image's statement, and each image is annotated with information about which objects it contains, their locations, sizes, and categories. And there are 250,000 pedestrians with keypoint annotations, this dataset focuses on scene interpretation, primarily derived from intricate daily scenes, with the objects within the images accurately marked by detailed segmentation for positioning.

4.2 Evaluation Indicators

To test the designed algorithm's effectiveness in recognizing multiple categories and detecting small targets, this paper examines its object classification and localization capabilities. We utilize mean Average Precision (mAP) [15] to verify the accuracy of the prediction results. This involves calculating the average accuracy (AP) for each category, which is then weighted and averaged to derive the mAP, a key metric for assessing the detection performance across multiple categories. Generally, a higher mAP indicates more robust detection capabilities of the model.

To calculate the mAP, the first step is to calculate the AP for each class, and the value of AP can be calculated by using Equation (3). where Loc_i denotes the prediction frame of the

model; Loc_i^{gt} denotes the true enclosing frame. IOU denotes the intersection and concurrency ratio of the predicted and true boxes.

$$AP = \frac{1}{N}\sum_{i=1}^{n} IoU(Loc_i, Loc_i^{gt}) \quad (3)$$

For a given test image, if the IOU exceeds the set threshold, the box that borders will be recognized as a correct example. The calculation is based on Equation (4).

$$IoU(Loc_i, Loc_i^{gt}) = \frac{Loc_i \cap Loc_i^{gt}}{Loc_i \cup Loc_i^{gt}} \quad (4)$$

AP is a measure of the quality of a class's detection, so mAP is a measure of the quality of multiple classes' detection. In multi class and multi target detection, after calculating the AP for each category, by the total number of categories, it is split. (5) displays the computation formula, in which C represents the aggregate amount of categories.

$$mAP = \frac{1}{C}\sum_{C}^{1} AP_c \quad (5)$$

4.3 Experimental Methods

The model's training and test sets are drawn from the COCO dataset. The original dataset is sampled for both the training and test sets, using a sample size ratio of 9:1. In the experiments, the gradient descent uses the entire training dataset in each iteration to compute the gradient, which involves randomly and evenly sampling multiple samples to form a small batch in each iteration round, and then using this small batch to calculate the gradient, a hundred cycles are iterated using the upgraded model. In order to hasten the model's convergence in the latter stages of training, the learning rate is dynamically adjusted to decay, and the initial value of 0.01 is taught to decay to 0.0001 in the later stages, and the experimental setup is displayed in Table 1.

4.4 Analysis and Outcomes of the Experiment

To obtain the target image, the image will be fused with the semantic details retrieved by U-Net and the convolutional features extracted by Faster R-CNN. Table 2 displays the experimental findings with the COCO dataset serving as the model's training and testing sets.

The accuracy of detecting is found to be greatly enhanced after the practice of combining the features extracted from the Faster R-CNN network with the semantic features extracted from the U-Net network. The algorithm improves the accuracy by an average of 11.9% at the IOU threshold of 0.5, and by 10.7% at an IOU threshold of 0.75. These values correspond to the IOU thresholds of 0.5 and 0.75 between this paper's method and previous methods, respectively. It is evident from the data that the algorithm presented in this research performs better at various thresholds.

Table 1. Experimental Setup

parameter category	Parameter name	parameter value
software environment	Ubuntu	18.04
	Python	3.6
hardware environment	GPU	Tesla T4
model training	Epoch	100
	Sample size ratio	9:1
	learning rate	0.01

Table 2. Experimental Results of Each Algorithm on COCO Dataset

Algorithm name	$IoU_{0.5}$ /mAP(%)	$IoU_{0.75}$ /mAP(%)
Faster R-CNN	45.3	23.5
Ours	57.2	34.2

In order to make the experimental data more complete, a different algorithm for removing redundant detection frames was added to the baseline model and the improved model, and Table 3 displays the experimental outcomes.

According to Table 3's results, the soft NMS approach for redundancy removal outperforms the NMS algorithm regards average accuracy at both the 0.5 and 0.75 IoU thresholds. This article uses the soft NMS technique to eliminate redundancy in the observed frames following the suggested object detection strategy that combines semantic segmentation networks. There is an average 4.9% improvement in accuracy at an IoU threshold of 0.5 and an average 1.8% improvement at an IoU threshold of 0.75. They have all improved, despite the fact that average accuracy varies significantly between IoU levels. The experimental results show that the final accuracy of this study is better than the original model.

Table 3. Comparison of Experimental Data

Algorithm name	$IOU_{0.5}$ /mAP(%)	$IOU_{0.75}$ /mAP(%)
Faster R-CNN+NMS	71.8	67.2
Faster R-CNN+Soft-NMS	73.3	68.3

Our+NMS	75.8	69.2
Our+Soft-NMS	**78.2**	**70.1**

The experimental findings were tested after the COCO data set was used to train and save settings. According to the experimental findings displayed in Figure 6, the algorithm was able to detect the target object even in situations where there were many objects and a disorganized background. Figures (a) through (b) demonstrate that even if the subject moves during the detection process, the target object may still be identified.

(a). The First Test Results

(b). The Second Test Result

(c). Third Test Result

(d). Fourth Test Result

Figure 6. Graph of Experimental Results

In the result diagram (a) - (b) of Figure.7, the size of the target object is very small, and our improved model can also detect the target well under different angles. This is enough to show that the target algorithm with semantic features is very feasible based on the classical algorithm Faster-R-CNN.

(a). The First Test Results

(b). The Second Test Result

(c). Third Test Result

(d). Fourth Test Result

Figure 7. Graph of Experimental Results

In the result diagram (a) - (b) of Figure.7, the size of the target object is very small, and our improved model can also detect the target well under different angles. This is enough to show that the target algorithm with semantic features is very feasible based on the classical algorithm Faster-R-CNN.

5. Conclusion

This article's primary contribution is the addition of the semantic segmentation network U-Net to the traditional Faster R-CNN object detection technique. This addresses the issue of imprecise identification brought on by intricate backdrop or inadequate feature values recovered when the target is too tiny in the original network. This algorithm combines the convolutional features extracted by Faster R-CNN algorithm with the semantic features extracted by the semantic segmentation network to increase feature expression ability and reduce false detection problems. The COCO dataset was used for training, and by comparing the mAP values that correlate to various IOU values, it was shown that this model is more effective at identifying smaller target items from photos with crowded backgrounds, the algorithm is then expanded to include the NMS and Soft-NMS algorithms, which eliminate redundant detection boxes and improve detection accuracy, the findings indicate when the U-Net network retrieves semantic features along with the Faster R-CNN algorithm, the accuracy of target object detection

increases.

Acknowledgements

This study was approved by the Foundation of the Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (Grant No.IPBED17), Graduate Innovation Program Project of Yan'an University, China (Grant No. YCX2023026).

References

[1] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, et al. A survey of modern deep learning based object detection models. Digital Signal Processing, 2022, 126: 103514.

[2] Shayhan Ameen Chowdhury, Mir Md. Saki Kowsar, Kaushik Deb. Human detection utilizing adaptive background mixture models and improved histogram of oriented gradients. ICT Express, 2018, 4 (4):216-220.

[3] Girshick Ross, Donahue Jeff, Darrell Trevor, Malik Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[4] Girshick, R..Fast R-CNN (Conference Paper).Proceedings of the IEEE International Conference on Computer Vision, 2015,Vol.2015: 1440-1448.

[5] Ren Shaoqing, He Kaiming, Girshick Ross, Sun Jian. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

[6] Linxiu Wu, Houjie Li, Jianjun He, Xuan Chen. Traffic sign detection method based on Faster R-CNN//Journal of Physics: Conference Series, 2019, 1176(3): 032045.

[7] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, Bo Ma. A Review of Yolo Algorithm Developments. Procedia Computer Science, 2022, 199: 1066-1073.

[8] Chenxi Zhang, Feng Kang, Yaxiong Wang. An Improved Apple Object Detection Method Based on Lightweight YOLOv4 in Complex Backgrounds. Remote Sensing, 2022, 14(17): 4150.

[9] Hakan Bilen, Andrea Vedaldi. Weakly supervised deep detection networks//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2846-2854.

[10] LiYang, Junhong Zhong, Yun Zhang, Sichang Bai, Guoshu Li, et al. An Improving Faster-RCNN with Multi-Attention ResNet for Small Target Detection in Intelligent Autonomous Transport with 6G. IEEE Transactions on Intelligent Transportation Systems. 2023, 24(7):P7717-7725.

[11] Jiao Lin, Dong Shifeng, Zhang Shengyu, Xie Chengjun, Wang Hongqiang. AF-RCNN: An anchor-free convolutional neural network for multi-categories agricultural pest detection. Computers and Electronics in Agriculture, 2020, 174: 105522.

[12] Hasib Zunair, A. Ben Hamza. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. Computers in Biology and Medicine, 2021, 136: 104699.

[13] Xu Zhaozhuo, Xu Xin, Wang Lei, Yang Rui, Pu Fangling. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. Remote Sensing, 2017, 9(12): 1312.

[14] Navaneeth Bodla, Bharat Singh, Rama Chellappa, Larry S. Davis. Soft-NMS--improving object detection with one line of code//Proceedings of the IEEE international conference on computer vision. 2017: 5562-5570.

[15] Prasanna Sheetal, ElSharkawy Mohamed. Improving Mean Average Precision (mAP) of Camera and Radar Fusion Network for Object Detection Using Radar Augmentation. Lecture Notes in Networks and Systems, 2023, 465:51-60.