Modified Object Detection Method Based on YOLO

Xia Zhao(✉), Yingting Ni, and Haihang Jia

School of Electronics and Information Engineering, Tongji University, Shanghai, China
zhaoxia@tongji.edu.cn

Abstract. YOLO (You Only Look Once), the 2D object detection method, is extremely fast since a single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. However, it makes more localization errors and its training velocity is relatively slow. Benefiting from the thoughts of cluster center in super-pixel segmentation and anchor box in Faster R-CNN, in this paper, we propose a modified method based on YOLO (shorted for M-YOLO). First, we substituted YOLOs last fully connected layer for a convolutional layer, on which the cluster boxes (some anchor boxes centered on cluster center) can completely cover the whole image at the beginning of training. As a result, the new structure can speed up the training process. Second, we increase the number of divided grids i.e. cluster centers, from 7×7 to the maximum 17×17, as well as the number of predicted bounding boxes, i.e. anchor boxes, from 2 to the maximum 9 for each grid cell. The measure can improve the IOU performance. Simultaneously, we also put forward a new kind of NMS (non-max suppression) to solve the problem aroused by M-YOLO. The experimental results show that M-YOLO improves the localization accuracy by about 10%, the convergence speed of the training process is also improved.

Keywords: Deep learning · Object detection · Cluster center
Anchor box

1 Introduction

Convolution Neural Networks (CNNs) had been widely used in 1990s, but then fell out of fashion with the rise of Support Vector Machines etc. [1]. In 2012, Hinton et al. won the first place on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by CNN. Their success led to a new wave of research, and more and more researchers use CNNs to improve the performance of image recognition and object detection.

In [1], Girshick et al. propose R-CNN (Region proposal Convolution Neural Networks). The method uses Selective Search to first generate around 2000 region proposals for the input image and computes features for each proposal using a CNN. Then the category-specific linear SVMs are employed to classify each

© Springer Nature Singapore Pte Ltd. 2017
J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 233–244, 2017.
https://doi.org/10.1007/978-981-10-7305-2_21

region. After classification, post-processing is used to refine the bounding boxes and eliminate redundant detections. Compared with the traditional detection methods R-CNN achieves excellent object detection accuracy on PASCAL VOC 2007 - about 58.5%. But it has a notable drawback that detection is slow C needing 47s for each image, because R-CNN performs a ConvNet forward pass for each region proposal, without sharing computation.

For fixing the disadvantages of R-CNN, [2] put forward the Fast R-CNN. The net-work takes a whole image and multiple regions of interest (RoIs) as input. The method first computes a convolutional feature map for the entire input image using the CNN. Then, each object proposal, the RoI pooling layer extracts a fixed-length feature vector from the feature map. Finally, each feature vector is fed into fully connected layers (FCs), which has two output vectors: softmax probabilities and per-class bounding-box regression offsets. Compared with R-CNN, the algorithm only takes once feature extraction for an input image, thus increasing the detection speed. At runtime, the network processes images in 0.3s (excluding object proposal time) while achieving higher detection quality with mAP 66%.

Though Fast R-CNN has reduced the detection time, it still depends on region proposal algorithms to predict object bounds, which is the computational bottleneck in detection systems. In [3], Girshick et al. introduce a Region Proposal Network (RPN) that is used to predict object bounds. The RPN is a fully convolutional network sharing convolutional features with the detection network, so that the marginal cost for generating proposals is small. By merging RPN and Fast R-CNN into a single network, Ross proposes an end-to-end object detection framework, named Faster R-CNN. The approach comparatively improves detection speed, up to 7 frames per second (FPS), while achieving object detection accuracy with mAP 73.2%. However, in real-time detection tasks, the speed is still unable to meet requirements.

In [4], Redmon presents YOLO (You Only Look Once), a new method to object detection. The system models object detection as a regression problem. It divides the image into 7×7 grids and for each grid cell predicts 2 bounding boxes, confidence for those boxes and 20 class probabilities. Since the method unifies the separate components of object detection into a single neural network, the YOLO model processes images in real-time at 45 FPS with mAP 63.4%. However, instead of generating region proposals, YOLO directly predicts multiple bounding boxes from the input image, which makes more localization errors and slow convergence speed.

SSD (Single Shot MultiBox Detector) is the current state-of-the-art object detection system [5]. The approach evaluates a set of default boxes of different aspect ratios at each location in several feature maps with different scales. At prediction time, the network generates category scores and box offsets for each default box by using small convolutional filters. Since it eliminates bounding box proposals and the subsequent pixel or feature resampling stage, SSD is much faster than methods that utilize proposal step and has comparable accuracy (58 FPS with mAP 72.1% on VOC2007 test, vs Faster R-CNN 7 FPS with mAP 73.2%). But SSD still performs poorly, when detecting small objects.

In the practice application, many scenes contain small objects to be detected and the intersection over union (IOU) between the predicted box and the ground truth is very important for some capture operations. Considering these factors and the characteristics of YOLO, we will have some modifications on YOLO and the new approach is named M-YOLO. First, combining the thought of the cluster center in super-pixel image segmentation with the thought of the anchor box in Faster R-CNN, M-YOLO generates the cluster boxes which can completely cover the whole image. Therefore, there is a smaller gap between predicted box and ground truth box than that of YOLO at the beginning of training. Second, the method substitutes YOLOs last fully connected layer for a convolutional layer, which generates category scores and box offsets for each cluster box. The new structure can speed up the training process. At the same time, the C-NMS (class-based non-max suppression) is designed, in order to solve the problem that the same object is identified as different categories. Compared with YOLO, M-YOLO improves the location accuracy and convergence speed, while keeping the detection accuracy.

The main contents of this paper are as follows: In Sect. 2, we briefly introduce YOLO as well as its problems. In Sect. 3, the new object detection approach M-YOLO is described in detailed. The performances of YOLO and M-YOLO are compared in Sect. 4.

2 YOLO

The detection process of YOLO is shown in Fig. 1. First of all, the input image is resized to 448×448. Then runs a single convolutional network on the image, bounding boxes as well as its confidence scores will be obtained. Finally, the NMS (non-max suppression) is used to threshold the resulting detections by the models confidence, the final class probabilities and bounding box coordinates are obtained as shown in the most right of Fig. 1.

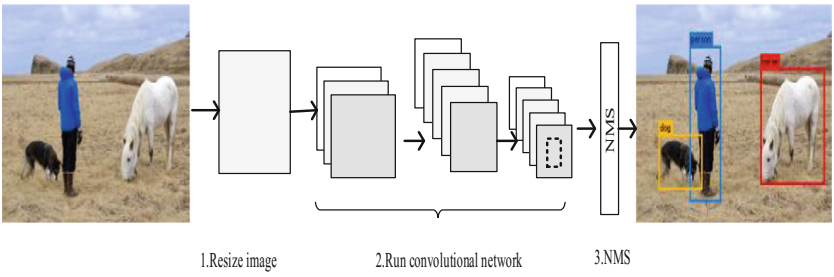

Fig. 1. The detection process of YOLO

YOLO divides the input image into 7×7 grids. Each grid cell corresponds to 2 bounding boxes. Each grid cell predicts one set of class probabilities (including

20 classes) regardless of the number of boxes. And each bounding box consists of 5 predictions: x, y, w, h and confidence. The (x, y) coordinates represent the center of the box relative to the bounds of the grid cell; the (w, h) represent the width and height of the box relative to the whole image. Thus, each grid has $(4 + 1) \times 2 + 20 = 30$ predictions and whole image has $7 \times 7 \times 30 = 1470$ predictions.

YOLOs CNN includes 25 convolutional layers, 4 pooling layers, 1 dropout layer, 1 fully connected layer and 1 detection layer. The dimension of the fully connected layer is 1470, denoting the position and confidence of the predicted bounding boxes and its class probabilities. For randomly initialized CNN, the output of the fully connected layer is also random at the beginning of the training process, i.e. there is a large gap between predicted box and ground truth box in the beginning. The sketch shown in Fig. 2, reveals the mapping between the fully connected layer and predicted box on the image. It needs a lot of iterations to get close to the ground truth box, resulting in slow convergence speed.

On the other hand, YOLO proposes far fewer bounding boxes, only $7 \times 7 \times 2 = 98$ per image, so its localization error is big.

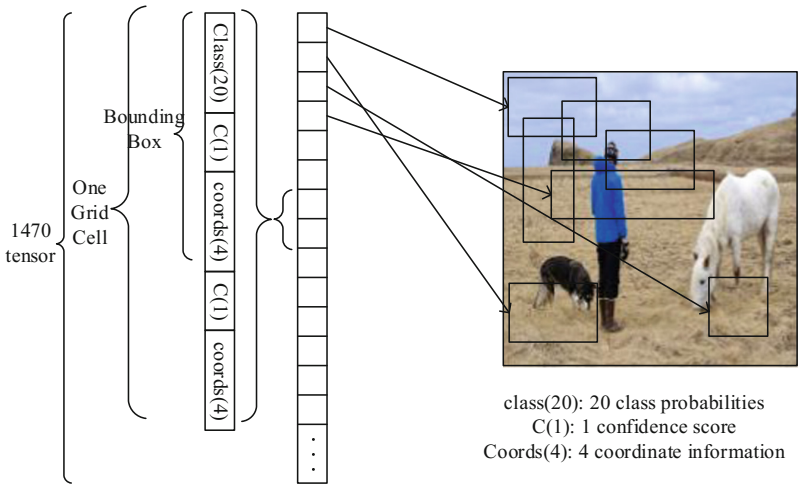

Fig. 2. Mapping between the fully connected layer and predicted box of YOLO

3 M-YOLO

3.1 The Principle of M-YOLO

In this paper, we propose a modified object detection method based on YOLO, i.e. M-YOLO, to improve the location accuracy and convergence speed. Benefiting from the thoughts of cluster center in the super-pixel segmentation and anchor box in Faster R-CNN, we introduce novel bounding boxes, called cluster

boxes, that serve as references at multiple scales and aspect ratios. Simultaneously, the new approach uses a convolutional layer to replace YOLOs fully connected layer.

The Anchor Box is first introduced in the Faster R-CNN [3]. Each feature point in the last convolution layer of RPN is as an anchor. And for each anchor, 9 kinds of anchor boxes can be pre-extracted by using 3 different scales and 3 different aspect ratios (Fig. 3). Compared with YOLO's 2 predicted bounding boxes, the anchor boxes of Faster R-CNN take into account the objects with different scales and aspect ratios. Therefore, on the basis of YOLO, we prepared to increase the number of predicted borders at multiple scales and aspect ratios, in order to improve the location accuracy.

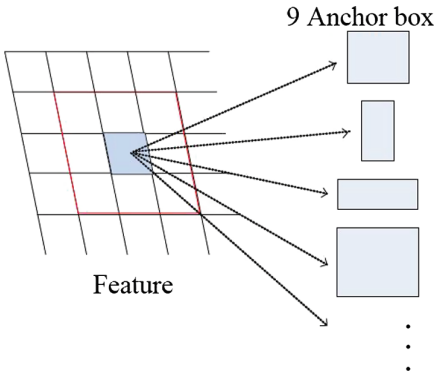

Fig. 3. Anchor boxes of the Faster R-CNN

Additionally, if the anchors in Faster R-CNN are mapped to the original image, they are some evenly distributed feature points in the image. Inspired by this, we managed to directly select some evenly distributed points from the input image, and take them as the center of predicted borders.

Super-pixel refers to an irregular pixel block with a certain visual representation, which consists of adjacent pixels with similar characteristics such as texture, color, brightness and so on [6]. In SLIC-based super-pixel segmentation, the first step is to select orderly and evenly distributed points in the input image as the cluster centers. The cluster centers can be fine-tuned to avoid appearing on the boundary of different color blocks. Using cluster centers as the initial centers of the iterative algorithm, the super-pixel with cluster center can be obtained after several iterations (Fig. 4). Due to the character of cluster centers, our method directly uses them as the center of predicted borders.

In this paper, our approach selects $n \times n$ cluster centers on the original image and predicts m bounding boxes with multiple scales and aspect ratios at each clustering center. The scale of the bounding boxes can be 1×1, 2×2, 4×4, the aspect ratios can be 1:1, 1:2, 2:1. The sketch map of cluster boxes is shown in Fig. 5, where $n = 7$, $m = 9$. The number of cluster centers and predicted boxes is important parameters affecting the detection results. We increase the number

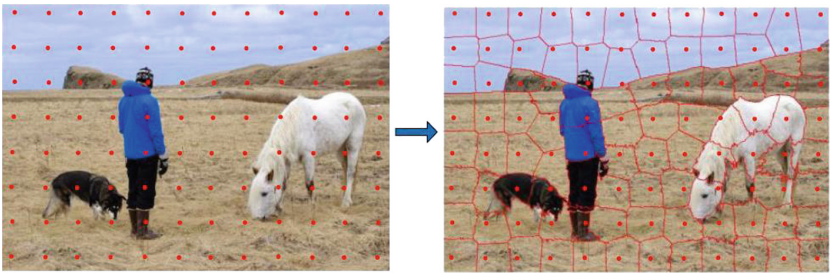

Fig. 4. Cluster centers and Super-pixel segmentation (Color figure online)

of cluster centers, from 7×7 to the maximum 17×17, as well as the number of predicted bounding boxes, from 2 to the maximum 9 for each grid cell. The measure can improve the IOU performance.

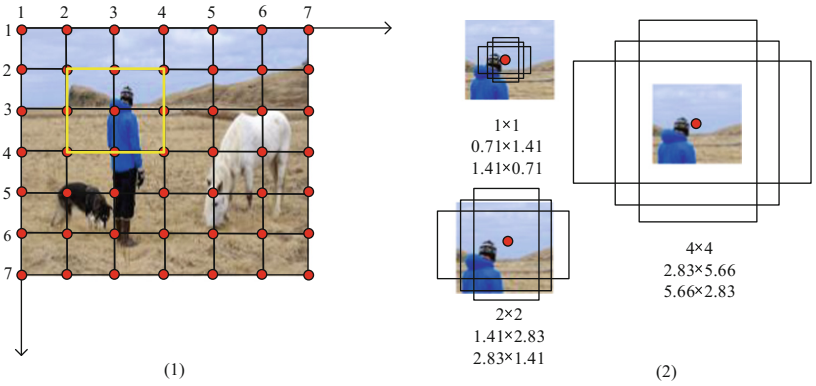

Fig. 5. The sketch map of cluster boxes

On bounding box predictions, each grid cell of YOLO shares 20 classification information since it only predicts two boxes and can only have one class. This constraint limits the number of nearby objects that the model can predict. In our method, all the boxes of each cluster center do not share 20 classification information, in order to detect nearby different objects.

M-YOLO uses the cluster boxes as the references and substitutes YOLOs last fully connected layer for a convolutional layer, on which the cluster boxes have stable sequences. The network generates category scores and box offsets for each cluster box. Since cluster boxes are orderly and evenly distributed in the input image, the gap between predicted box and ground truth box is as small as possible at the beginning of the training process. Thus this model performs well when trained and benefits convergence speed. The sketch mapping between the new conventional layer and predicted box on the image is shown in Fig. 6.

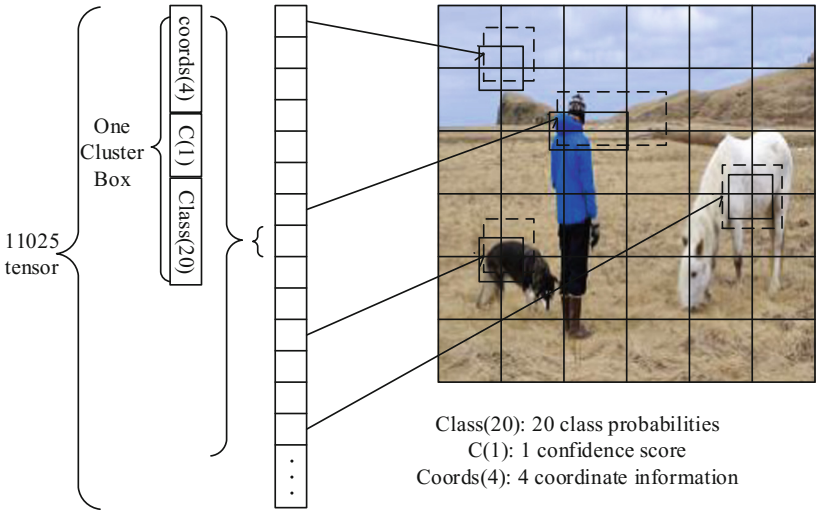

Fig. 6. Mapping between the new conventional layer and predicted box of M-YOLO

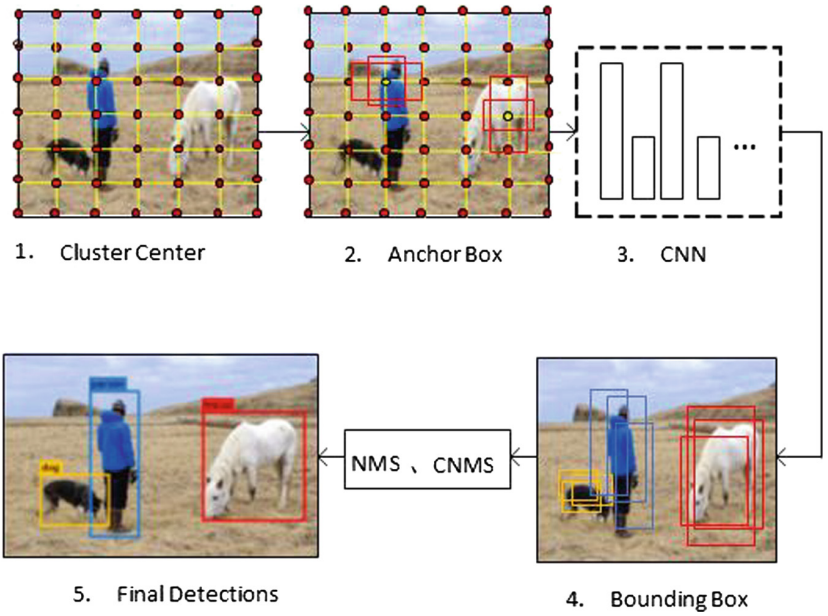

Fig. 7. The object detection process of M-YOLO

The object detection process of M-YOLO is shown in Fig. 7. It selects $n \times n$ clustering centers on the original image and predicts m bounding boxes with multiple scales and aspect ratios at each clustering center. Then, the image is input to the CNN to predict confidence, class probabilities and box offsets for each cluster box. Finally, use NMS to eliminate redundant detections.

3.2 C-NMS (Class Non-max Suppression)

In order to detect different objects in a wide range of scales and aspect ratios, the 9 bounding boxes of each cluster center do not share classification information. Thus, there is a problem that the same object is identified as different categories by adjacent bounding boxes. As shown in Fig. 8(1), the object 'horse' is identified as horse or cow. To solve this problem, we introduce the C-NMS. For all bounding boxes of adjacent cluster centers, if the class probabilities is greater than a certain probability score (we use 0.2) and the overlap degree between adjacent bounding boxes is greater than a threshold (we use 0.5), we think the same object is identified as different categories by redundant bounding boxes. According to experimental experience, we directly choose the bounding box having the largest area as the final object bounding box. Figure 8(2) is the result after using C-NMS.

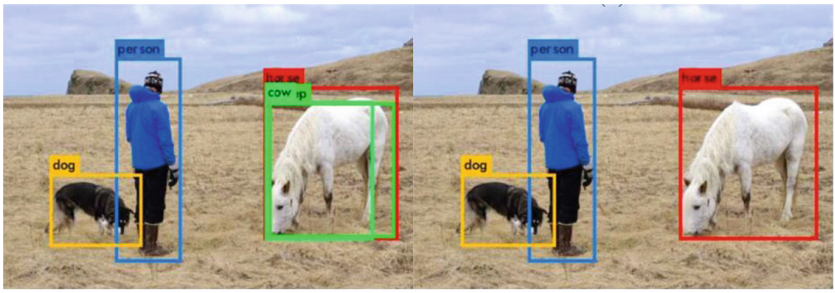

Fig. 8. The result after using C-NMS

3.3 Training

In this paper, the experimental environment is Ubuntu and we use the Darknet framework for all training and testing. We train the network on the VOC2007train + VOC2007val + VOC2012train + VOC2012val dataset, which consist of about 15k images over 20 object categories. Throughout training we use a batch size of 64, a momentum of 0.9 and a decay of 0.0005.

Learning rate is a key parameter in training the network. If the learning rate is too high, the weights may diverge optimal values; and if the learning rate is too small, it takes long time to find optimal values. Figure 9 is the change of learning rate for M-YOLO and the LR is 0.0005. According to the increase in iteration times, learning rate is constantly changing. To prevent model divergence caused by unstable gradients, we start at a relatively small learning rate. Then we slowly raise the rate from LR to 10LR. We continue training with 10LR for 14000 iterations, then LR for 10000 iterations, and finally 0.1LR for 10000 iterations.

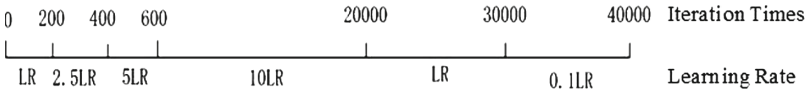

Fig. 9. The change of learning rate

4 Experimental Results

In our model, we need to determine two parameters: the number of cluster centers (n) and that of bounding boxes for each cluster center (m). The number of cluster centers changes from 7×7 to 17×17. The bounding boxes are set according to the setting rules of anchor boxes in Faster R-CNN. Simultaneously, in view of YOLO's 2 boxes and Faster R-CNNs 9 boxes, we also set the number of boxes to 5 for comparison. We comprehensively evaluate it on the VOC2007 detection dataset, consisting of about 5k test images. Performance metrics include the mean average precision (mAP), recall rate (R), the intersection over union (IOU). We also provide results on the VOC 2007 for YOLO.

As we can see from Table 1, the model M-YOLO-15-5 ($m = 15$, $n = 5$) scores 61.07% mPA, which is the best of all models and slightly higher than YOLO. The model M-YOLO-17-9 ($m = 17$, $n = 9$) has highest IOU (79.99%). Compared with YOLO, it improves IOU about 10% and has more accurate object bounding boxes. Additionally, M-YOLO-17-9 also pushes recall rate to 89.95%, which is more 13% than that of YOLO.

Figure 10(1) is the curve of mAP with the number of cluster centers n and the bounding box number m. It can be seen that mAP reaches the peak, when $n = 15$, $m = 5$. Figure 10(2) shows the curve of IOU with the number of cluster

Table 1. PASCAL VOC2007 test detection results

Method	Boxes	Cluster center	mAP	IOU	Recall
YOLO	2	7	60.36	69.66	76.55
M-YOLO	5	7	46.86	63.59	68.85
		9	54.15	72.20	80.47
		11	57.18	74.46	82.52
		13	60.40	76.78	85.58
		15	61.07	77.88	87.56
		17	60.74	78.71	87.93
	9	7	47.45	66.77	69.48
		9	55.31	74.87	81.72
		11	58.43	77.04	84.50
		13	60.54	78.65	87.11
		15	60.77	79.65	88.88
		17	60.25	79.99	89.53

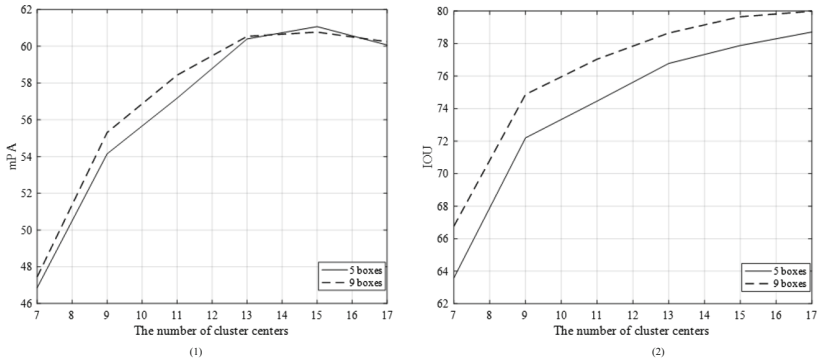

Fig. 10. The curve of performance metrics with the number of cluster centers (n) and the border number (m)

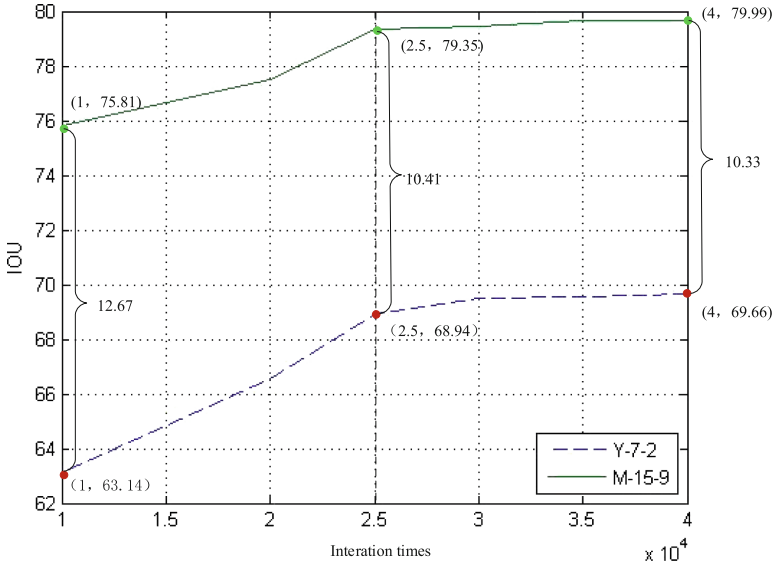

Fig. 11. The IOU values of different iteration times

centers n and the border number m. We can see that the IOU tends to be stable as the bounding box number increasing, and the IOU value of 9 bounding boxes is higher than that of 5 bounding boxes as a whole. Figure 11 shows the IOU values of different iteration times. When the iteration times are 10,000, the IOU of M-YOLO-15-5 is 12.67% higher than that of the YOLO. However, when the iteration times are 40,000, the IOU of M-YOLO-15-5 is 10.33% higher than that of the YOLO. It can be seen that the model M-YOLO-15-9 accelerates the convergence speed. Figure 12 shows selected examples of object detection results on the VOC 2007 test set using the M-YOLO system.

Fig. 12. Examples of object detection results of M-YOLO

5 Conclusion

In this paper, based on the YOLO, we propose an improved object detection method M-YOLO, which utilizes the cluster center in the super-pixel segmentation and the Anchor Box of Faster RCNN. The experimental result confirms that M-YOLO improves the accuracy of object bounding boxes by about 10% and the recall rate, while keeping the detection accuracy.

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition, pp. 580–587 (2014)
2. Girshick, R.: Fast R-CNN. In: International Conference on Computer Vision, pp,. 1440–1448 (2015)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)

4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Computer Vision and Pattern Recognition, pp. 779–788 (2016)
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: single shot multibox detector. ECCV **1**(2016), 21–37 (2016)
6. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)