Research on YOLOv3 algorithm based on darknet framework

Rong Guo¹, Shixin Li², Kun Wang³
¹Tianjin University of Technology and Education, Tianjin, China

Corresponding author and e-mail: Rong Guo, guorong949160@163.com

Abstract. Traffic jams and accidents occur frequently in modern cities. In the context of smart cities, intelligent transportation can be effectively controlled through target detection technology. In view of the problems of slow detection speed and low accuracy of traditional vehicle detection algorithms, a YOLOv3 algorithm based on K-means ++ is proposed. The accuracy of bounding box detection is improved by the K-means ++ algorithm. Compared with the traditional YOLOv3 detection algorithm, the improved algorithm improves the detection speed and accuracy. Experiments show that the improved algorithm has a higher recognition rate for small targets in the actual test, while reducing the false detection rate and improving the accuracy of the algorithm.

1. Introduction
With the development of modern cities and the improvement of people's consumption level, urban traffic jams and traffic accidents occur frequently, which causes great inconvenience to people's production and life. In response to such problems, the concept of smart city has been proposed, hoping to control intelligent transportation, so as to effectively solve traffic congestion and provide protection for illegal traffic evidence [1]. At the same time, for the application research of target detection technology, domestic and foreign scholars have gradually carried out research work on vehicle detection. In the field of traditional machine learning, scholars extract target features and input the extracted features to support vector machines, iterators, and other classifiers for classification and recognition, but the disadvantage of this method is that the scale of the data is not large, and the generalization ability is relatively small and poor. It is difficult to accurately identify the target [2]. In the field of deep learning, the application of convolutional neural networks combines artificial networks and convolution operations, resulting in many deep learning target detection algorithms. YOLO (You Only Look Once) algorithm is a brand-new algorithm for end-to-end detection, which can maintain a high accuracy rate while performing rapid detection [3]. The single-stage detection algorithm YOLOv3 has become the mainstream detection algorithm in engineering applications due to its good detection accuracy and speed. This paper improves the YOLOv3 algorithm. Compared with the traditional YOLOv3 detection algorithm, experiments show that this algorithm improve the accuracy of target detection.

2. YOLOv3 algorithm
2.1. The basic idea of YOLO algorithm
The basic principle of YOLOv1 algorithm is to divide an image into SxS grid cells. If the center of an object falls in this grid, this grid is responsible for predicting this object [4]. There are B bounding
boxes in each grid. Each bounding box has 5 parameters: $x$, $y$, $w$, $h$, confidence. The first four parameters are used to record the position and size of the bounding box. Confidence represents accuracy of the object contained in the predicted box, and its calculation formula is:

$$
Pr(\text{Object}) \cdot \text{IOU}_{\text{pred}} \cdot \text{IOU}_{\text{truth}}
$$

(1)

If there is an object in a grid cell, the first item will be 1, otherwise it will be 0. The second term is the IOU value between the predicted bounding box and the actual target box.

Each grid cell also predicts a category information, that is, the probability value judged as a certain category, which is recorded as category $C$. In the YOLO model, there are 20 categories, so there are 20 parameters in $C$. From the above, the output form of each image is: $S \cdot S \cdot (S \cdot B + C)$ vector. Its network structure is shown in Figure 1:

Figure 1. YOLOv1 algorithm network structure diagram.

The YOLOv1 algorithm uses a convolutional network to extract features, and then uses a fully connected layer to obtain predicted values. The detection layer of YOLOv1 algorithm contains 24 convolutional layers and 2 fully connected layers.

Although the YOLOv1 algorithm has the advantages of fast operation and accurate recognition, the algorithm still has some defects. For example, if there is no object in the grid cell, then $Pr(\text{object}) = 0$, if there is an object, then $Pr(\text{object}) = 1$. Although there are two Bounding Boxes in a grid cell, they share the same set of classification probabilities, so the same cell can only recognize the same object, which will cause inaccurate object recognition and missed detection.

2.2. YOLOv3 algorithm

The YOLOv3 algorithm introduces the anchor box in the Faster-RCNN algorithm, which can effectively improve the accuracy of the bounding box to identify objects [5]. The anchor box is the shape and size of several boxes that are counted from all ground truth boxes in the training set and most frequently appear in the training set, using the k-means clustering method [6].

YOLOv3 uses the method of YOLO9000 to predict the bounding box, and determines the anchor box through size clustering. Four predictions for each bounding box network, with a coordinate offset of $t_x$, $t_y$, $t_w$, $t_h$. If a cell of the feature map is offset by the coordinate $(c_x, c_y)$ in the upper left corner of the picture, and the pre-selection box of the bounding box’s size is $p_w$, $p_h$, that is, the size of the anchor, then the generated prediction coordinate is $b_x$, $b_y$, $b_w$, $b_h$, which is the feature map level. And $g_x$, $g_y$, $g_w$, $g_h$ is the true value mapping on the feature map. By predicting the offset $t_x$, $t_y$, $t_w$, $t_h$ to make $b_x$, $b_y$, $b_w$, $b_h$ and $g_x$, $g_y$, $g_w$, $g_h$ consistent, as shown in Figure 2 specifically.
3. Improve YOLOv3 algorithm

For the YOLOv3 algorithm, the original K-means algorithm is used to randomly select K points in the data set as the clustering center. By adding statistical a priori (or human) experience to the model, multiple experiments are required during the experiment and with uncertainty. This paper proposes a YOLOv3 algorithm based on K-means ++[7-9].

The K-means ++ algorithm selects K clustering centers according to the following idea: assuming that n initial clustering centers \((0 < n < K)\) have been selected, when selecting the \(n + 1\) clustering center: points further away from the current \(n\) cluster centers will have a higher probability of being selected as the \(n + 1\) cluster center. When selecting the first cluster center \((n = 1)\), the random method is also adopted[10].

The following are the specific steps for applying the K-means ++ algorithm to the YOLOv3 algorithm:

1. The original clustering data used is a detection data set with only labeled boxes. YOLOv3 will generate a TXT file containing the location and category of the labeled boxes, where each line contains \((x_j, y_j, w_j, h_j)\), \(j \in \{1, 2, \ldots, N\}\), that is, ground truth boxes are the coordinates of the figure, \((x_j, y_j)\) is the center point of the frame, \((w_j, h_j)\) is the width and height of the frame, and \(N\) is the number of all labeling boxes;

2. First, \(k\) cluster center points \((W_i, H_i), i \in \{1, 2, \ldots, k\}\) are given, where \(W_i, H_i\) are the width and height dimensions of the anchor boxes. Since the position of the anchor boxes is not fixed, there is no coordinate of \((x, y)\), only the width and height;

3. Calculate the distance \(d = 1 - IOU(\text{labeling box, cluster center point})\) between each labeling box and each cluster center point. The center point of each labeled box coincides with the clustering center during calculation, so that the IOU value can be calculated, namely \(d = 1 - IOU((x_j, y_j, w_j, h_j), (x_i, y_i, W_i, H_i)), j \in \{1, 2, \ldots, N\}, i \in \{1, 2, \ldots, k\}\). Assign the labeling boxes to the clustering center closest to "distance";

4. After all the labeling boxes are assigned, the cluster center point is recalculated for each cluster. The calculation method is \(W'_i = \frac{1}{N_i} \sum w_j, H'_i = \frac{1}{N_i} \sum h_j\), and \(N_i\) is the number of labeling boxes of the \(i\) cluster. It is to average the width and height of all the labeled boxes in the cluster;

5. Repeat steps 3 and 4 until the cluster center changes little.

4. Experiment and result analysis
The configuration of the experimental platform is: the operating system is Windows7, the CUDA version is 10.0, the CUDNN version is 7.4.1, and Tensorflow2.0 is used for training.

During the verification process, The match is successful when the prediction probability is greater than 50% and the prediction box matches the actual marked rectangular box IOU is greater than 0.5. The actual test is a car in a video taken on a road. Test 1, 25, 50, 75 and 100 frames, the experimental results are shown in the figure below.

![Frame 1](image1)
![Frame 25](image2)
![Frame 50](image3)
![Frame 75](image4)
![Frame 100](image5)

**Figure 3.** Test results of frames 1, 25, 50, 75 and 100. (the left side is the yolov3 algorithm result, the right side is the algorithm result of this article)

The test accuracy rate is shown in Table 1:

| Algorithm          | frame 1 | frame 25 | frame 50 | frame 75 | frame 100 |
|--------------------|---------|----------|----------|----------|-----------|
| Yolov3 algorithm   | 98.2%   | 92%      | 80.4%    | 70.5%    | 90%       |
| Algorithm of this paper | 100%    | 95%      | 85.3%    | 75%      | 98%       |

| Table 1. Comparison of the accuracy of the two algorithms. |

Experimentally tested at frame 1 and 100, this algorithm has better recognition ability and accuracy, and the recognition rate has been significantly improved. The 25, 50, and 75 frames of this algorithm are more fuzzy than YOLOv3 algorithm. Small targets perform better, and the YOLOv3 algorithm
misses detection. In summary, the algorithm in this paper performs better and more accurately when identifying targets.

5. Conclusions
This article is based on the YOLOv3 algorithm. By improving the clustering algorithm of the preselected bounding box size, the K-means in the original YOLOv3 algorithm is changed to the K-means ++ algorithm, which can effectively improve the accuracy of the bounding box. Experimental results show that: compared with the original network of YOLOv3, the improved YOLOv3 algorithm has high detection accuracy, low misrecognition rate and good robustness for small targets in actual road target detection tasks.

Acknowledgement
Fund Project: Tianjin Science and Technology Commissioner Project (19JCTPJC54800); Tianjin University of Technology and Education Graduate Innovation Fund (YC19-30).

References
[1] BIAN SHANFENG, ZHANG QINGHUI. Real-time vehicle detection algorithm based on improved YOLO v2 [J]. Electronic Quality, 2019, 10: 19-22. (In Chinese)
[2] HAN ZHUO. Design and implementation of target detection model on TensorFlow platform [D]. Beijing: Beijing University of Posts and Telecommunications, 2019. (In Chinese)
[3] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. 2015.
[4] CAI CHENGTAO, WU KEJUN, LIU QIUFEI, et al. Real-time panoramic multi-target detection based on improved YOLO algorithm [J]. Computer Engineering and Design, 2018, 39 (10): 3259-3271. (In Chinese)
[5] HONG SONG, GAO DINGGUO, THREE ROWS ONLY. The vehicle and pedestrian detection method based on YOLO_v3 [J]. Computer Knowledge and Technology, 2020, 16 (8): 192-198. (In Chinese)
[6] Joseph Redmon, Santosh Divvala. You only look once: Unified, real-time object detection [C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas: IEEE, 2016: 779-788.
[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [C]. Computer Vision and Pattern Recognition, 2016, 770-778.
[8] ZHOU XIAOYAN, WANG KE, LI LINGYAN. Survey of target detection algorithms based on deep learning [J]. Electronic Measurement Technology, 2017, 40 (11): 89-93. (In Chinese)
[9] CHANG LIANG, DENG XIAOMING, ZHOU MINGQUAN, et al. Convolutional Neural Networks in Image Understanding [J]. Journal of Automation. 2016, 42 (09): 1300-1312. (In Chinese)
[10] DING SONGTAO, QU SHIRU. Detection of interest area of traffic target based on deep learning [J]. Journal of China Highway, 2018, 31 (9): 167-174. (In Chinese)