Research on Application of Improved YOLO V3 Algorithm in Road Target Detection

JIN Zhao-zhao¹, ZHENG Yu-fu*¹

¹ School of Electronics and Information Engineering. Lanzhou Jiaotong University, Lanzhou, Gansu, 730070, China

Corresponding author’s e-mail: 1217602826@qq.com

Abstract: Due to the lack of average accuracy and missed detection in the process of real road scene target detection through YOLO V3 network, the improvement scheme is put forward. The K-means clustering algorithm is used to replace the K-means clustering algorithm in the original network to analyze the anchor number and aspect ratio of the Udacity data set, in order to make the obtained parameters more suitable; in addition, in order to improve the performance of the road target detection algorithm, the existing network output is upgraded, and a 104 × 104 feature detection layer is added, and the feature map output by 8 times sampling can be output by 2 times up sampling, and 4 the feature maps of down-sampling are stitched together, and the 104 × 104 feature maps obtained can effectively reduce the disappearance of features. Through the experimental results, we can see that compared with the improved YOLOV3 algorithm, the average detection accuracy of the improved algorithm for road target detection is quite high to 3.17%, and the missing detection rate is reduced by 5.62%.

1. Introduction
In the field of intelligent driving, road target detection technology embodies an extremely critical role, and is widely used in vehicle assisted driving systems [1]. In recent years, with the improvement of computer hardware, artificial intelligence technology and the continuous improvement of different types of target detection technologies, research on road target detection methods based on deep learning is of great significance to the field of intelligent driving [2].

Because the current computer and deep learning technology are in a rapid development stage [3], and deep learning has various target detection algorithms [4-5] and target tracking algorithms [6], it is widely used in road target detection related problems. Based on the convolutional neural network in deep learning, the focus is on the initiative of learning the target data set, which can better improve the model. In the current deep learning aspect and widely used target detection algorithms can be divided into two categories: On the one hand, it is the One-stage target detection algorithm. This algorithm does not require a regional candidate network (RPN) module, and can use the network to directly generate relevant information such as target location and category, and include regression problems in the target frame. The positioning problem in, the algorithm belongs to end-to-end target detection. Representative algorithms include SSD (single shot Multibox detector)[7], YOLO (you only look once)[8], YOLO 9000[9], YOLO V2[10], YOLO V3[11], etc. On the other hand, is the Two-stage target detection algorithm. Through this algorithm, target detection can be divided into two stages. The first stage first calculates the sample candidate frame, and the second stage is mainly for the detection
network. Under its conditions, the location and category of the detection target candidate area are classified and identified. The representative network can be divided into several types: R-CNN [12], Fast R-CNN [13], Faster R-CNN [14] and Mask R-CNN [15]. Since single-step target detection algorithms can extract target information without the need for an RPN module, the network can have directness in acquiring target location and category information. It can be seen that this type of algorithm is better in terms of detection speed; However, due to the lack of RPN module, its detection accuracy is not high. The YOLO V3 type is a single-step target detection network. In addition to its faster detection speed, it also uses the anchor idea contained in the Faster R-CNN network to enhance its target detection rate to a certain extent, but the YOLO V3 network is in road scenes.

In order to make the recall rate and accuracy rate of YOLO V3 network during road target detection is improved. First, the K-means++ algorithm [16] with less randomness is used instead of the K-means algorithm [17] to perform clustering analysis on the road target data set to determine the number of clusters and the aspect ratio, and modify the anchor parameters in the YOLO V3 algorithm accordingly. Secondly, in view of the detection accuracy of road targets and the characteristics of small detection targets in the far field of view, the improved network structure of the YOLO V3 algorithm can enhance its accuracy and recall during the detection process. In the simulation experiment, by using Udacity (Udacity self-Driving) data, the vehicle in the dataset is used as the detection target, and the network structure in YOLO V3 is improved based on this premise. First of all, adjust the second residual module in the Darknet-53 structure to help extract the feature information with more completeness, and in the current state of the second residual module, add 2 additional residual units. Subsequently, in order to fully integrate the semantic information in the high-level, the 8 times down-sampling output feature map in the Darknet-53 structure was subjected to 2 times up-sampling processing, and then the 2 times up-sampling output feature map and the second residual the difference module outputs feature maps for feature splicing. Finally, in order to fully prevent the YOLO V3 network from disappearing features after the improvement, it is necessary to replace the improved detection target output layer, and replace the original 6 convolution units with 2 convolution units and 2 residuals unit. Compare the difference between the adjustment of the YOLO V3 network, and conduct a simulation test on it. The results show that the target detection is performed after the optimization and improvement of the YOLO V3 network. The recall rate and detection rate, average detection accuracy, missed detection rate have different degrees of enhancement.

2. YOLO V3 algorithm
The network structure of the YOLO V3 model still uses the Darknet-53 network model, as shown in Fig. 1. Darknet-53 includes 5 residual modules and 53 convolutional layers. Its main idea refers to the residual neural network [18].

Darknet-53 alternately uses the module network of ‘convolution→ convolution→ residual’ to form a fully convolutional network, which propagates forward through the network, keeping the size of the feature map synchronized with the step size of the residual module. The Darknet-53 network has performed 5 down-sampling of the image, each of which has a sampling step size of 2 and a maximum step size of 32; Then the image is subjected to 32-fold down-sampling, 16-fold down-sampling, and 8-fold down-sampling processing to obtain 3 target detections with scale differences, Then the image is subjected to 32-fold down-sampling, 16-fold down-sampling, and 8-fold down-sampling processing to obtain 3 target detections with scale differences, the feature maps are respectively $13 \times 13$, $26 \times 26$, $52 \times 52$; The first feature map is suitable for detecting large targets, the second feature map tends to detect medium targets, and the last one is mostly used for small targets. Finally, the output feature map is sampled and feature fusion is performed, and finally the detection task of the target is completed.
3. YOLO V3 algorithm improvement

The article takes the road object detection problem as the main research object, based on the Udacity data set. For the Udacity data set, the previous network definition anchor boxes and network hierarchical structure are used as the road object detection research object the applicability is poor. Therefore, the K-means++ clustering algorithm is used to carry out the cluster analysis of the road targets in the data set, and then the actual situation of the road target detection is carried out pertinent analysis, so as to adjust the hierarchical structure of the network.

3.1. Cluster analysis of the target frame of the data set

The YOLO V3 algorithm is based on the idea of Faster R-CNN anchor frame, and clusters the data set through a clustering algorithm to obtain the parameter information of the road anchor frame. YOLO V3 original candidate frame anchor frame parameters are clustered through the COCO data set, and 9 is the number of anchor frames, including prediction scales that are different from three. Since the data set used for road target recognition in this article is the Udacity data set, the original anchor parameters are not suitable for the target recognition in this article. K-means algorithm and K-means++ algorithm are often used in the cluster analysis of anchor frame size for target detection. Different k values are obtained by comparing the two clusters, so as to select the clustering algorithm.

As shown in Fig. 2, it can be seen that when the value of k is in an increasing trend, the objective function values obtained by the K-means algorithm and the K-means++ algorithm both show a decreasing trend, and tend to be smooth, and the clustering effect is gradually improved. However, in the process of reducing the objective function and improving the clustering effect, the function curves of the two algorithms have stronger smoothness and more stable trends. Therefore, the K-means++ algorithm reduces the clustering deviation to a certain extent, so the K-means++ algorithm is the most suitable for cluster analysis and processing of data sets.
Fig. 2 The objective function corresponding to different k values

By using the intersection ratio (IoU) as the standard for the measurement of cluster analysis, the gradually increasing k is used as the independent variable, and the intersection ratio of the dependent variable realizes the cluster analysis of the data set. The objective function f of clustering can be expressed as:

$$f = \arg \max \sum_{i=1}^{n} \sum_{j=1}^{k} I_{\text{IoU}}(b,c)$$

In formula (1), b represents the sample; the cluster center is represented by c; the number of samples at the k-th cluster center is represented by nk; the total number of samples is represented by n; the number of clusters is represented by k; the cluster center frame and the cluster The intersection and union ratio between the class boxes is expressed by passing; i is the sample number; j is the sample number of the cluster center. Finally, 9 anchors are determined, and the size of the anchor frame as shown in Table 1 is obtained through the K-means++ clustering algorithm. Through the analysis of the results in Table 1, it can be seen that when the value of k is close to 9, larger clusters will appear to generate redundancy, which will be the result of the improvement of the clustering parameter value over k. Compared with the original parameters of the YOLO V3 algorithm, the clustering result is significantly smaller than the original width and height parameters.

Table 1 Real-world frame aggregation results of Udacity dataset

| K=7   | K=8   | K=9   | K=10  | K=11  |
|-------|-------|-------|-------|-------|
| (12,9)| (12,9)| (11,9)| (11,9)| (9,7) |
| (21,14)| (28,12)| (19,13)| (15,17)| (17,11)|
| (38,23)| (34,21)| (36,22)| (37,22)| (33,19)|
| (57,41)| (55,39)| (56,35)| (55,37)| (57,35)|
3.2. Improve network structure

When the target is detected and processed through the YOLO V3 network, the focus is on the detection of the sampled output, using 8 times, 16 times, and 32 times feature maps as the completion target. In addition, the feature map with location information limitation due to 8 times down-sampling output includes. It can be seen that the detection rate can be substantially improved due to the increase in the amount of feature information obtained by the network to obtain the target to be detected. At the same time, feature extraction of the target to be detected is a 52×52 feature map generated by the 4-fold down-sampling output in the original network. At the same time, information about the newly added unit 2 in the lower layer can also be obtained to understand the specific positioning of the second stub network. In the YOLO V3 network, the output is a 16 times down-sampling feature map of 26×26, and 2 times upsampling is used on the basis of the original feature map, so that the feature map. Subsequently, the feature image generated by combining the second residual block in Darknet with the double output is combined with the feature information and the feature map is obtained by 4 times output down-sampling based on the feature information. The four output feature maps of 13×13, 26×26, 52×52, 104×104 are sequentially fused to obtain a fusion target detection layer with new features, thereby realizing the detection of road targets.

In order to improve the recall rate and detection accuracy of the network during road target detection, based on the K-means++ algorithm, the clustering results are obtained through the Udacity data set, the clustering results are obtained through the Udacity data set, and it is guaranteed that when the YOLO V3 network performs three-scale output detection, it can be combined with the new feature target detection layer to achieve the detection of road targets and improve the detection performance of the network. Before the output layer is generated by detecting the target through the YOLO V3 network, it has 6 DBL units and a single 1×1 convolution. You can refer to the core idea of the DSSD (deconvolutional single shot detector) network [19]. In order to prevent the disappearance of the gradient, and at the same time strengthen the characteristics of its reuse ability, the 6 DBL units are replaced with 2 DBL units and residual units, which can improve the detection efficiency of the network and avoid the phenomenon of feature disappearance. The improved YOLO V3 network structure is shown in Fig.3.
Since the YOLO V3 network will have zooming or cropping when accurately measuring the feature image, no matter what the resolution of the input detection image is, it will eventually be zoomed or cropped to an image with a resolution of 416×416. In the experiment, after processing the Udacity data set, two sub-data sets are obtained. There are 9423 and 15000 in the pictures, and the resolution of the data set is 1920×1200. In addition, because the initial picture has a higher resolution, you can adjust the resolution through the batch conversion function provided by IrfanView, resulting in a compressed picture of the same name with a resolution of 416×416.

4. Simulation comparison analysis

The YOLO V3 algorithm for target detection is based on multiple different scales, and as the most representative end-to-end target detection algorithm at the moment, it also has the function of detecting various types of targets, and is aimed at the detection of small targets. Its detection performance is better. Focus on comparing the performance changes of YOLO V3 target detection algorithm before and after improvement.

4.1. Network performance test

The test set in the data set is tested on the trained and improved YOLO V3 network and YOLO V3 network, and the recall rate and detection accuracy rate of the main detection target are calculated for the test process. Table 2 compares the recall rate, accuracy rate and missed detection rate of the two target detection algorithms, YOLO V3 and the improved YOLOV3.

Table 2 Comparison of accuracy, recall and missed detection rates of target detection algorithms in this article

|                  | Detection rate (%) | Recall rate (%) | Missing detection rate (%) |
|------------------|--------------------|----------------|----------------------------|
| YOLO V3          | 84.3               | 87             | 15.31                      |

![Improved YOLO V3 network structure diagram](image)
Experimental results show that the accuracy rate of the YOLO V3 network before and after the improvement has increased by 2.4%, and the recall rate has increased by 5%, and the missed detection rate has been reduced by 5.62%. In order to further detect the network performance, the accuracy (AP) obtained by the road target detection is calculated for the two networks, and the value depends on the recall rate and the accuracy rate to measure the accuracy, which can be the most intuitive evaluation the accuracy standard of the detection model is used to analyze and study the actual effect of target detection. As shown in Figure 4, the AP detection results obtained before and after the improvement of YOLO V3, the improved network improves the accuracy of target detection from 74.54% to 80.71%.

![Fig.4 AP curves of the two networks. (a) YOLOV3 network AP curve; (b) Improved YOLOV3 network AP curve](image)

Compare the memory of related data before and after the improvement of YOLO V3, divide it into 6 types of target detection, and calculate the average accuracy average mAP produced by the detection of different targets before and after the improvement. And the lower the value, the poorer overall performance of various target detections performed by the model. As shown in Table 3, it shows the results obtained after the two networks are tested. After the improvement of the YOLO V3 network, the average accuracy of target detection obtained by mAP is 82%, which is obviously on the basis of 78.83% of the original algorithm has seen an increase.

|                  | AP/%  | mAP/% |
|------------------|-------|-------|
| Car              | 81    |       |
| Motorbike        | 81    |       |
| Train            | 80    |       |
| Bicycle          | 79    |       |
| Bus              | 77    |       |
| Person           | 75    | 78.83 |
| Improved YOLO V3| 84    | 82    |
|                  | 84    | 82    |
|                  | 83    | 82    |
|                  | 82    | 80    |
|                  | 79    | 79    |
|                  | 82    | 82    |

4.2. Result analysis
In this paper, K-means ++ clustering algorithm, the cluster re-anchor the parameters obtained are compared with parameters of the original anchor YOLO V3 algorithm, re-concentrated and the clustering result is significantly less than the original width and height parameters, which reduces the
clustering deviation to a certain extent. The YOLO V3 network improves the down-sampling feature target detection layer that increases the output by 4 times on the original basis, so that in the process of road target detection, its accuracy rate is 2.4% higher than before, and the recall rate is also increased by 5.0%. The average accuracy value generated during target detection has been improved by 3.17% before the comparison and improvement. At the same time, the problem of the original high missed detection rate in the road scene is fully improved. It can be seen that after the improvement of the YOLO V3 network, the detection effectiveness of the scene target has been greatly improved, and the missed detection rate has been reduced by 7.13%, reducing the occurrence of missed detections. The road target detection after the improvement of the YOLO V3 network is significantly stronger than the unimproved YOLO V3 network in terms of detection performance.

5. Conclusion
To apply the improved YOLO V3 network to the road target detection scene, firstly, the K-Means++ clustering algorithm must be used to re-cluster the size and number of anchor boxes in the Udacity dataset to ensure that the anchor parameters suitable for road target detection are obtained. Because there are many types of targets to be detected in the road target detection scene, the height and width of various targets are not unique, and the horizontal feature expression of the target is less than the vertical feature expression. To a certain extent, increase the horizontal feature expression of the feature map and help improve the network the detection performance. In order to prevent the feature from disappearing due to the network depth being too large, the image can be sampled three times first, and then the output feature map can be up-sampled twice, and then stitched with the drawing after the two down-sampling processing, and finally get feature target layer with 4 times down-sampling size; In order to improve the feature multiplexing level of the network, 2 residual units and 2 DBL units are used to replace the 6 DBL units in the original network to improve the feature multiplexing capability of the network. After the experiment, I learned that after optimization and adjustment of the YOLO V3 algorithm, the accuracy, recall, and average precision of the original YOLO V3 algorithm will be greatly improved when detecting targets.

References
[1] Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. [J]. Acta Automatica Sinica, 2017, 43(08):1289–1305.
[2] Yin Hong-Peng, Chen Bo, Chai Yi, Liu Zhao-Dong. Vision-based object detection and tracking: a review.[J]. Acta Automatica Sinica, 2016, 42(10): 1466–1489.
[3] Wang D C, Chen X N, Zhao F, et al. Vehicle detection algorithm based on convolutional neural network and RGB-D images[J]. Laser Optoelectronics Progress, 2019, 56(18): 181003.
[4] HE Kaiming, ZHANG Xianyu, REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence 2014,37(09):1904-1916.
[5] Wang Wenxiu, Fu Yutian, Dong Feng, Li Feng. Infrared ship target detection method based on deep convolutional neural network [J]. Journal of Optics, 2018, 38 (07): 160-166.
[6] Luo Haibo, Xu Lingyun, Hui Bin, Chang Zheng. Research status and prospects of target tracking methods based on deep learning [J]. Infrared and Laser Engineering, 2017, 46 (05): 14-20.
[7] Liu W, Anguevok D, Erhan D, et al. SSD: single shot multibox detector[C]/European Conference on Computer Vision. 2016:21-37.
[8] REDMON J, DIVVALA S, GIRSHICK R, et al.You only look once: unified, real-time object detection [C]/Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C, USA: IEEE Computer Society: 2016:779-788.
[9] REDMON J, FARHADI A. YOLO 9000: better, faster, stronger. [EB/OL]. Ithaca, NewYork: Cornell University Library, 2016[20171101]. https://arxiv.org/abs/1612.08242.
[10] Li Yunpeng, Hou Lingyan, Wang Chao. Vehicle target detection in complex scenes based on YOLOv2 [J]. Television Technology, 2018, 42 (5): 100-106.
[11] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[12] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014: 8-13
[13] Girshik R. Fast R-CNN [C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.
[14] REN S, HK, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transaction son Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
[15] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99): 1-1.
[16] Huang G, Liu Z, Weinberger KQ, et al. Densely connected convolutional networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4700-4708.
[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]// Computer Vision and Pattern Recognition. 2016: 777-778.
[18] Zhu Wang, Jalil Boukhobza, Zili Shao. Preserving SSD lifetime in deep learning applications with delta snapshots[J]. Journal of Parallel and Distributed Computing, 2019, 133.
[19] Multimedia; New Findings on Multimedia Described by Investigators at Hefei University of Technology (A real-time recognition method of static gesture based on DSSD) [J]. Journal of Technology, 2020.