International Journal of New Developments in Engineering and Society
ISSN 2522-3488 Vol. 6, Issue 3: 38-42, DOI: 10.25236/IJNDES.2022.060307

Car, Cyclist and Pedestrian Object Detection Based on YOLOv5

Yuwei Xiang, Qianxiao Fei

SHIEN-MING WU School of Intelligent Engineering, South China University of Technology, Guangzhou, Guangdong, 510000, China

Abstract: *The target detection of car, cyclist and pedestrian is of great significance to the realization of automatic monitoring and artificial intelligence assisted driving system. In order to rapidly detect car, cyclist and pedestrian, You Only Look Once (YOLO) model is applied to car, cyclist and pedestrian detection tasks in this paper. Experimental result shows that among the four versions of YOLOv5, the mean average precision (MAP) is more than 91.7%, and YOLOv5x has the best recognition ability, which reaches 92.6%. Among the three categories, car recognition accuracy is the highest, followed by cyclist and pedestrian.*

Keywords: *Target detection; Car; Cyclist and Pedestrian; YOLOv5*

1. Introduction

With the progress of science and technology and the development of the industry, autonomous driving has become one of the key research tasks in the automotive field. Object detection system is an important part of automatic safe driving detection technology. It is an important premise to ensure the safe driving of automatic driving to detect the objects on the road efficiently and accurately.

Traditional object detection methods can be roughly divided into three parts: region selection, feature extraction (SIFT[1], HOG[2], etc.) and classifier. Their main problems are as follows: on the one hand, the sliding window selection strategy is not targeted and time complexity is high; on the other hand, their feature robustness is poor. Deep learning object detection can greatly improve these problems. Deep learning object detection is mainly divided into two-stage object detection and single-stage object detection. The accuracy of two-stage object detection is high, but the number of frames processed per second is low, which cannot meet the real-time detection function required by automatic driving technology. Single-stage object detection directly transforms the problem into a regression problem and can get the final detection directly, so it can achieve the function of real-time detection. Deep learning object detection includes R-CNN[3], YOLO[4], SSD[5], etc. We choose YOLO for object detection.

The core idea of YOLOv1 is to unify the object detection into a regression problem related to position and category. There is no process of obtaining region proposal, and the position and category of Bounding Box are directly regression at the output layer. On the basis of maintaining the processing speed of YOLOv1, YOLOv2[6] introduces batch normalization, uses prior frames and clustering to extract information of prior frames to achieve Better, Faster and Stronger. YOLOv3[7] introduces the residual structure and uses the convolutional layer with step 2 to replace the pooling operation for down-sampling, realizing the full convolutional neural network and adopting Feature Pyramid Networks for multi-scale feature fusion. Based on YOLOv3, YOLOv4[8] uses the multi-scale feature fusion Network of Spatial Pyramid Pooling and Path Aggregation Network, uses Mish activation function, and uses Mosaic data enhancement, DropBlock regularization and other improved strategies in the input side. YOLOv5 is an innovative model based on YOLOv3 and YOLOv4 models, which is characterized by small model and fast speed and is suitable for mobile terminals. We use the YOLOv5 series models to process the data.

YOLO has been studied in road target detection. Shi et al. [9] used improved YOLOv2 network to achieve an average detection accuracy of 87.84% for different types of vehicles. Qi et al. [10] achieved a vehicle detection accuracy of 88.6% by using improved YOLOv3 network. Peng et al. [11] obtained an accuracy of 88.9% by using YOLOv4Tiny to detect vehicles, and 92.1% by using YOLOv4 to detect vehicles. WY Hsu and WY Lin [12] obtained an accuracy of 91.6% by using adaptive fusion of multi-scale YOLO.

Although the deep learning algorithm has achieved good results in road target detection tasks, it still

ISSN 2522-3488 Vol. 6, Issue 3: 38-42, DOI: 10.25236/IJNDES.2022.060307

has some aspects to be improved: 1) It is difficult to adapt to complex background and multi-target scenes. 2) The real-time pedestrian detection effect is not good. In order to ensure real-time detection, YOLOv5, which has the most average accuracy and frames per second, is used to establish a road target detection model.

2. The principle of YOLOv5

Figure 1 shows the overall block diagram of YOLOv5s object detection algorithm. For an object detection algorithm, we can usually divide it into four general modules, including input, backbone, neck network and head output, corresponding to the four modules in the figure above.

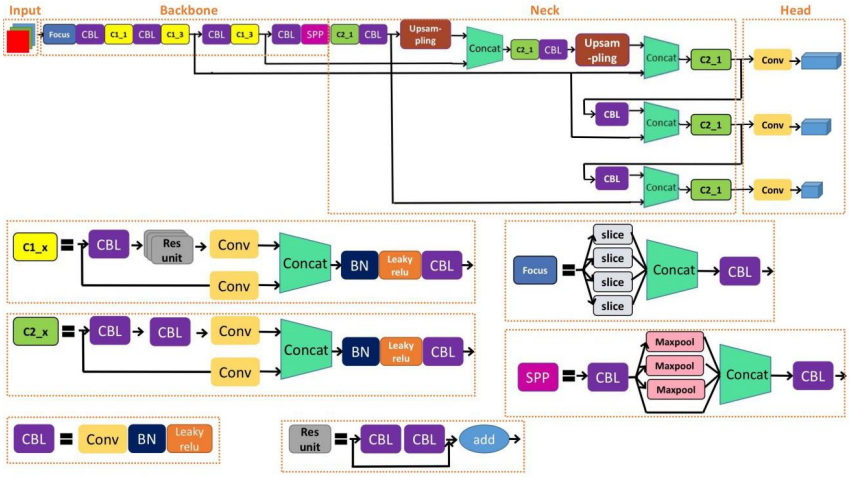

Figure 1: The architecture of the YOLOv5s

Input: This stage usually includes image pre-processing, which will scale the input image to the input size of the network and carry out normalization and other operations. In the network training stage, YOLOv5 uses Mosaic data enhancement operation to improve the training speed and accuracy. A method of adaptive anchor frame calculation and adaptive image scaling is proposed.

Backbone: The backbone is usually the network of some excellent classifier. This module is used to extract some general feature representation. In YOLOv5, both CSPDarknet53 structure and Focus structure are used.

Neck network: Neck network is usually located in the middle of the backbone network and the head, which can further improve the diversity and robustness of features.

Head output: Head is used to complete the output of object detection results. For different detection algorithms, the number of output branches is different, usually including a classification branch and a regression branch.

YOLOv5 is a single-stage object detection algorithm. Based on YOLOv4, this algorithm adds some new ideas to improve its speed and accuracy greatly.

The specific improvement methods are as follows:

1) At the input end of model training stage, some improvement ideas are proposed, including Mosaic data enhancement, adaptive anchor frame calculation and adaptive image scaling.

2) The backbone integrates some new ideas in other detection algorithms, such as Focus structure and CSP structure.

3) FPN+PAN structure is inserted between BackBone and Head output.

4) The loss function GIOU_Loss during training and DIOU_nms during prediction box screening are improved.

According to the different depth of network structure, YOLOv5 is divided into YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The training parameters of different YOLOv5 are shown in Table 1.

ISSN 2522-3488 Vol. 6, Issue 3: 38-42, DOI: 10.25236/IJNDES.2022.060307

Table 1: Training parameters of different YOLOv5

Model	Picture Size	Number of Training	Depth_multiple	Width_multiple
YOLOv5s	1242x375	300	0.33	0.50
YOLOv5m	1242x375	300	0.67	0.75
YOLOv5l	1242x375	300	1.0	1.0
YOLOv5x	1242x375	300	1.33	1.25

3. Experimental results and analysis

3.1. Data sources

Data sets from Raw Data of KITTI, which contains 7481 road images. Since the data set is abundant, we do not have to worry about fitting. When driving on autopilot in an emergency, we not only need to quickly and accurately identify the object, but also need to choose a safe route according to different situations. We divide the vehicles and people on the road into three categories: Car, Cyclist, Pedestrian. Because the average speed of each class is different, distinguishing the three categories helps the autonomous system choose the most appropriate path. The pictures containing the object are manually marked by software, and PASCAL VOC format is obtained after processing. Finally, the obtained data set was divided into training set and testing set in a 4:1 ratio (5928 images in the training set and 1553 images in the testing set).

3.2. Model evaluation index

Average precision (AP) and Mean Average Precision (mAP) were used to evaluate the model. Precision represents the proportion of predicted results that are correct:

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

Recall represents the proportion of all targets that were correctly predicted:

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

TP represents the number of cases correctly classified as positive, FP represents the number of cases incorrectly classified as positive, and FN represents the number of cases incorrectly classified as negative.

Curve integral $AP = \int_0^1 p(r)dr$ is obtained by precision-recall (PR) for each category, and mAP is obtained by averaging AP values of all categories. mAP@0.5 indicates that when IOU is set to 0.5, the average mAP of all categories.

3.3. Experimental environment

The experiment uses Pytorch as the software framework. The hardware environment of model training is Intel(R) UHD Graphics 620 and NVIDIA GeForce MX130. Accelerate Graphics Processing Unit (GPU) based on Compute Unified Device Architecture (CUDA) to improve computer Graphics computing capability.

4. Results analysis

Figure 2 shows the training results of YOLOv5s on the data set. It can be seen from the figure that the precision decreases with the increase of recall. Based on other training results, we can obtain the prediction accuracy and average accuracy of car, cyclist and pedestrian.

ISSN 2522-3488 Vol. 6, Issue 3: 38-42, DOI: 10.25236/IJNDES.2022.060307

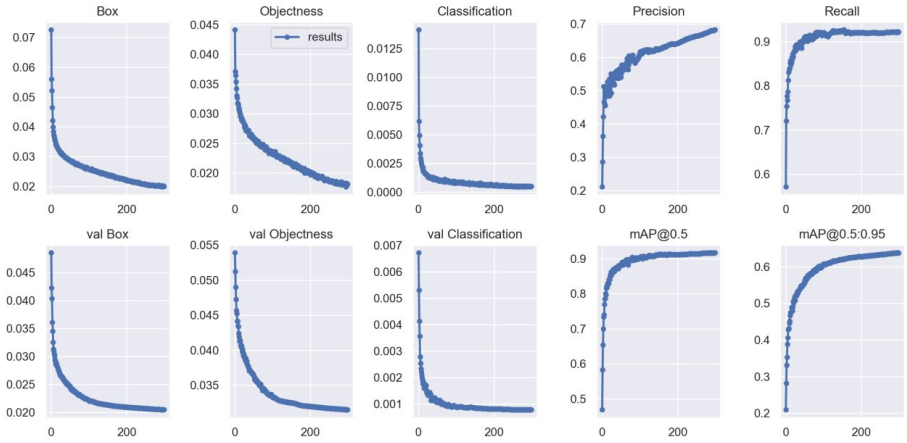

Figure 2: The training results of YOLOv5s

As shown in the Table 2, the mAP increases as the network depth and width increase. On the whole, the YOLOv5x method has the highest accuracy, while the YOLOv5s method has the lowest accuracy. In the same model, the accuracy of car is the highest, cyclist is the second, and pedestrian is the lowest.

Table 2: Comparison of detection model results

Model	Car	Cyclist	Pedestrian	mAP@.5
YOLOv5s	**0.982**	**0.911**	**0.858**	**0.917**
YOLOv5m	**0.982**	**0.914**	**0.863**	**0.920**
YOLOv5l	**0.983**	**0.920**	**0.868**	**0.924**
YOLOv5x	**0.983**	**0.923**	**0.871**	**0.926**

5. Conclusion

In this paper, object detection of car, cyclist and pedestrian is carried out based on YOLOv5 models. With the continuous expansion of the network, the network depth is deepening, and the number of convolution kernels is increasing, which means that the network width is further expanding. At the same time, the weight files generated after training are getting bigger and bigger, which indicates that more parameters are included, the accuracy of the whole model is further improved, and the detection speed will become slower and slower with the increase of the model. By comparing the results of the four models, it can be concluded that YOLOv5x has the highest accuracy and slowest detection speed among the four methods, while YOLOv5s has the lowest accuracy and fastest detection speed. Therefore, YOLOv5s can better meet the real-time demand and is more suitable for unmanned driving and other fields.

References

[1] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints [J]. International Journal of Computer Vision, 2004, 60(2):91-110.
[2] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE, 2005.
[3] Girshick R, Donahue J, Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Computer Society. IEEE Computer Society, 2013.
[4] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. IEEE, 2016.
[5] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector [J]. 2015.
[6] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
[7] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [J]. arXiv e-prints, 2018.
[8] Bochkovskiy A, Wang C Y, Liao H. YOLOv4: Optimal Speed and Accuracy of Object Detection [J]. 2020.
[9] Shi B, Li X, Nie T, et al. Multi-object Recognition Method Based on Improved YOLOv2 Model [J]. Information Technology and Control, 2021, 50(1):13-27.

ISSN 2522-3488 Vol. 6, Issue 3: 38-42, DOI: 10.25236/IJNDES.2022.060307

[10] Qi Y, Shi H, Li N, et al. Vehicle Detection Under Unmanned Aerial Vehicle Based on Improved YOLOv3[C]// 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019.
[11] Peng H, Guo S, Zuo X. A Vehicle Detection Method Based on YOLOV4 Model. 2021.
[12] Hsu W Y, Lin W Y. Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection [J]. IEEE Access, 2021, PP (99):1-1.