Research on driverless vehicle vision algorithm

Xinchao Liu[1,*], *Ying* Yan[1], and *Haiyun* Gan[1]

[1]Tianjin University of Technology and Education, Tianjin 300222, China

Keyword: YOLOv3-SPP3-tiny, Complex scenes, Target detection, Model pruning.

Abstract. Obstacle detection in complex urban traffic environment has become an important part of unmanned vehicle optimization, and its complexity brings great challenges to the reliability of unmanned target detection. YOLOv3 in deep learning algorithm has a good detection effect in target detection, but it has certain defects in detecting targets in complex urban traffic environment. In this paper, the spatial pyramid module is added to YOLOv3 to improve the extraction of data features of the deep model. Then, on the basis of optimized network, the target detection algorithm is streamlined by combining layer pruning and channel pruning. The streamlined algorithm is called YOLOv3-SPP3-Tiny. Comparing the experimental results of YOLOv3-SPP3-tiny and YOLOv3 on Street Scenes dataset, the Precision is improved by 2.77%, the average precision (mAP) is increased by 0.87%, the Total BFLOPS is reduced by 94.49%, and the Inference time is reduced by 80.39%. Experimental results show that the model YOLOv3-SPP3-tiny algorithm is more conducive to unmanned object detection in complex urban road environment.

1 Introduction

Target detection algorithm is one of the most basic research fields of driverless cars. With the rise of deep learning, a large number of target detection algorithms are applied to improve the detection accuracy. As a difficult point in the study of target detection, urban traffic environment is also an attractive research point. Especially for driverless vehicles, computer vision is one of its main research directions. Improving the accuracy and speed of target detection in urban traffic environment is crucial to reduce the accident rate. Traditional target detection algorithm on the precision and speed can achieve a satisfactory result, based on the deep learning algorithms in machine vision task (including target detection, classification and tracking) [1-6] made a major breakthrough, the machine vision approach from the target image parsing out the computer can understand information, machine vision the cognition of image mainly in four aspects: 1.Classification mainly analyzes the categories of the target;2.Positioning To determine the location of the target;3. Detection is the combination of classification and positioning to determine the object class

* Corresponding author: liuxinchao_com@sina.com

and its position;4.There are two types of segmentation main one is the semantic segmentation category is the instance segmentation, semantic segmentation will each pixel point in the image annotation for some categories of objects, while another instance segmentation for the combination of semantic segmentation and target detection, marked images are similar other different forms, but the machine with human perception or there is a big gap. Machine vision will be blocked by objects, low resolution, and bad weather, which will reduce the detection effect. In addition, the complexity of urban traffic environment from motor vehicles, pedestrians and non-motor vehicles to buildings, trees and sidewalks also leads to missed detection and false detection of target detection algorithm. In the actual detection of driverless vehicles in the face of complex traffic environment, there are mainly the following challenges:

(1) Influence of light: Most vehicles have the characteristics of high reflected light in strong light, and the detection effect in weak light will also reduce the feature extraction effect of target detection.

(2) Complex and diverse types: The complex types of vehicles and the changeable clothing of road pedestrians in the urban traffic environment have great demands on the performance of target detection.

(3) Target occlusion: Due to the large number of vehicles in urban traffic, the complex traffic flow and the changeable movement characteristics of road pedestrians make it difficult to timely capture effective feature information due to the sudden appearance of objects in the process of man-car, car-car intersection or turning.

(4) Background interference: Complex urban traffic contains a large number of billboards, various types of shop fronts, cloudy sky and other complex background information, which is a difficult problem to obtain target information.

In order to overcome the above obstacles, Street Scenes of urban traffic data sets containing nine object categories were first selected in this work. Secondly, an urban traffic detection model based on YOLOv3[7-9] is optimized. On the framework of YOLOv3, the spatial pyramid structure was added to the three detection heads to enhance the deep feature extraction effect, and then pruning strategy [10-12] was developed to reduce the redundancy of the network model and improve the detection efficiency. In this paper, YOLOv3-SPP3-Tiny is proposed based on the detection problem in urban traffic environment to improve the target detection of driverless vehicles in urban traffic.

2 Related work

2.1 Network optimization

After continuous improvement, YOLOv3 algorithm of high precision was optimized by YOLO [7-9] series algorithm. YOLOv3 algorithm was developed on Darknet, a light learning framework. It has the advantage of high speed and can make full use of multi-core processor and GPU parallel operation.YOLOv3 [7-9] in order to make the network model to achieve better detection effect, on the basis of the original to add a lot of good performance of the convolution layer of 3×3 and 1×1, at the same time enhanced shortcut links in a network, the number of structure of the improved algorithm of target detection YOLOv3 precision than YOLO [7-9] series of other target detection precision of the algorithm is much higher. However, in the complex urban traffic environment, the target detection of YOLOv3 will be subject to object occlusion, multi-motion state and bad weather, which will reduce the detection effect. In this paper, based on the YOLOv3 algorithm, a space pyramid module [13] (SPP) is added to improve the deep features. The space pyramid module uses proportional pooling, which can output the features of fixed

dimensions without considering the size of the extracted feature map, mainly replacing the full connection layer. The improvement point is to divide the image into several parts, while the traditional pooling is to specify the size of one part. Another improvement is that no matter how big the input is, the output size is the same, so at this time, the last layer of feature map is cropped in proportion and then thrown into the pool layer to output features, which can eliminate the problem of inconsistent size of the input image.

In order to fully enhance the feature extraction of deep structure in the experiment, the SPP module was added between the 5th and 6th convolutional layers of YOLOv3 to improve the detection effect, and YOLOv3-SPP3 was combined. As shown in Figure 1 below:

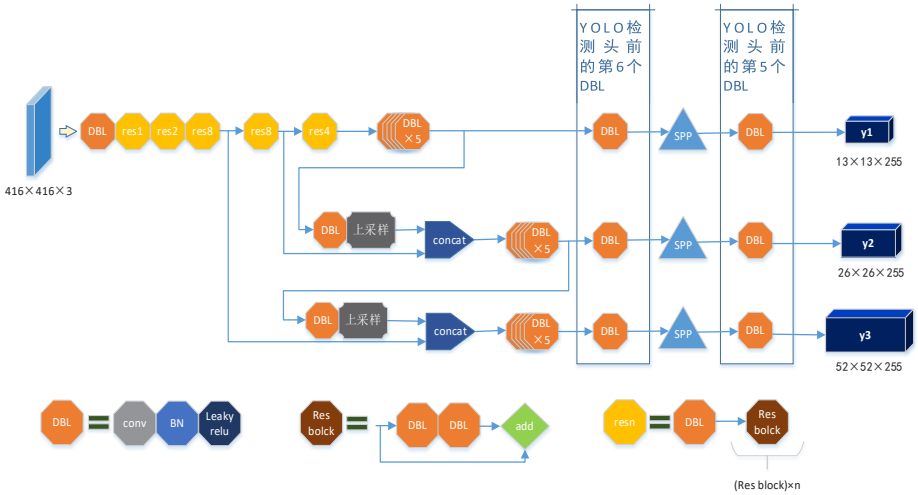

Fig. 1. Network structure diagram of YOLOv3-SPP3.

3.2 Network pruning

The redundant network part of the target detection algorithm is reduced by the combined pruning strategy [10-12] to improve the detection efficiency. Based on the optimization of network YOLOv3-SPP3, network model pruning is carried out. figure 2 pruning iteration process is used to extract the enhanced network YOLOv3-SPP3-Tiny.

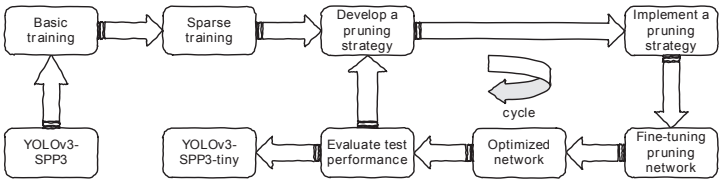

Fig. 2. Iterative process of YOLOv3-SPP3-tiny.

Pruning iteration process of YOLOv3-SPP3 network :(1) basic network training;(2) Network sparse training;(3) Evaluation of the index adjustment pruning method of the deep model;(4) Network fine-tuning of the algorithm model completed by pruning improves the detection effect of the model;(5) Evaluate the pruning algorithm model and then pruning deployment to achieve the optimal performance, that is, YOLOv3-SPP3-Tiny.

For the pruning method based on the above iterative method, firstly, the experiment improves channel level channel sparsity by adding L1 regularization to the channel scaling factor [14,15], which is conducive to network structured pruning. Then the pruning rate of the channel is adjusted according to the introduced global channel threshold and local channel security threshold. Finally, based on the planned channel pruning, layer pruning strategy was added to evaluate the convolutional layer associated with shortcut layer, and the convolutional layer with minimum scale factor was removed to improve detection efficiency. Then YOLOv3-SPP3-Tiny algorithm was optimized. This experiment is an improved strategy based on the method proposed by Liu[12], which is mainly improved as a coarse-grained deep model exploration [16] method to optimize the depth model detector with stronger real-time performance.

3 Experiment

In order to optimize the YOLOv3 target recognition algorithm to better meet the target detection in driverless urban traffic, based on the optimization of deep feature extraction, a combined pruning strategy was developed to improve the detection efficiency of the model.

3.1 Data set

In order to verify the effectiveness of the optimization algorithm in driverless urban traffic environment, a StreetScenes data set produced by StanlyBileschi was selected, which was filmed in and around Boston, Massachusetts, USA, with DSC-F717 camera. Its data set is manually tagged and contains nine object categories, namely pedestrians, cars, bicycles, trees, sky, buildings, sidewalks, roads and shops, all of which are strictly completed in the same tagged manner. The StreetScenes data set consists of one image, 1280 x 960 pixels, and is divided into training sets, verification sets, and test sets on an 8:1:1 scale for experiments.

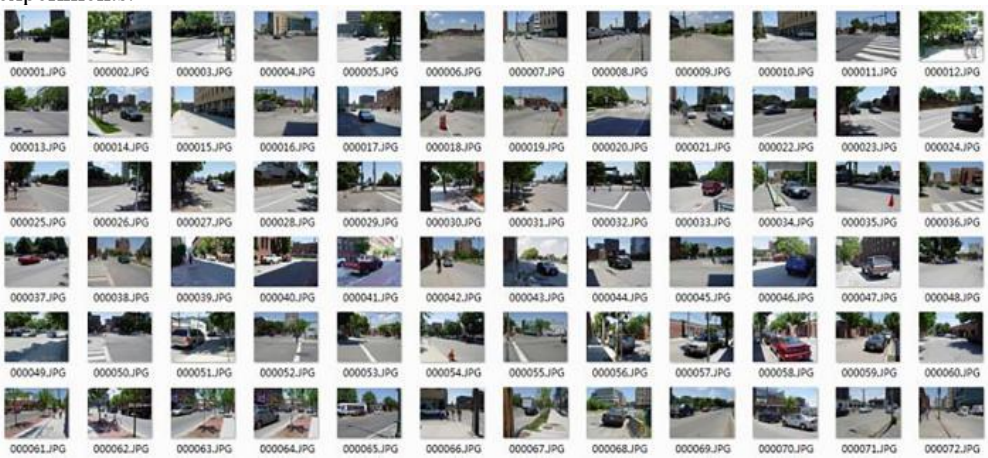

Fig. 3. Shows an example of a dataset.

3.2 Model training

The experiment was carried out in PyTorch, a deep learning framework. In the training setting, four data sets were loaded into the network model at one time to learn 100 EPOchs. The learning rate of the initial training is 0.01, and the learning rate will decline tenfold

when the network training reaches 70% and 90% of the total learning content. The weight attenuation is 0.001 and the momentum is 0.9.

Sparse training: When YOLOv3 had trained 100 epochs, it was done to speed up network pruning. Sparse training of 300 epochs followed by training of 100 epochs.

3.3 Performance indicators

The following indexes are used to analyze the performance of the optimized model: 1.Precision; 2.Recall (recall); 3.mAP (average detection accuracy); 4.Total BFLOPS (total floating point operation); 5.Inference time (inference time of each picture); 6.Parameters (parameter size); 7.Volume (model space).

4 Analysis of experimental results

The target detection model with the best detection effect is obtained by comparing the evaluation indexes of the basic model and different optimization models in the table (YOLOv3-SPPN-X,N = 1 or 3 represents the number of space pyramid pools, and X represents pruning X%).

Table 1. Experimental results of different optimization models.

Model	Input size	Precision	Recall	mAP	Total BFLOPS	Inference time(ms)	Parameters(M)	Volume(MB)
YOLOv3	416	83.30	73.00	76.10	67.29	29.00	60.52	241.40
	864	86.50	88.10	91.90	285.65	88.20		
YOLOv3-SPP1	416	86.20	82.70	82.40	67.67	30.70	62.67	251.60
	864	89.50	91.90	92.50	285.27	90.40		
YOLOv3-SPP3	416	86.90	82.60	84.90	73.18	35.96	64.85	257.80
	864	89.90	92.90	94.10	310.02	101.68		
YOLOv3-SPP1-95	864	88.60	88.60	90.20	17.10	19.00	0.70	2.90
YOLOv3-SPP1-96	864	87.50	85.40	87.10	15.20	16.80	0.56	2.20
YOLOv3-SPP1-97	864	85.50	77.20	81.50	11.99	15.90	0.35	1.40
YOLOv3-SPP3-95	864	89.70	89.50	93.60	19.19	20.50	0.79	3.30
YOLOv3-SPP3-95	***864***	***88.90***	***88.40***	***92.70***	***15.74***	***17.30***	***0.58***	***2.30***
YOLOv3-SPP3-96	864	86.50	78.90	89.70	12.57	16.50	0.39	1.50
YOLOv3-SPP3-97								

Table 1 shows the optimization results obtained in the experiment. First of all, as can be seen from the Input-size column in the table, the evaluation index of YOLOv3 also changed significantly when the network input image size was increased from 416×416 to 864×864. Precision increased by 3.2, Recall increased by 15.1, mAP increased by 15.8, Total BFLOPS increased by 218.36, and Inferencetime increased by 59.2ms.Therefore, the training model with network input image size of 864×864 was set to optimize different pruning rates.

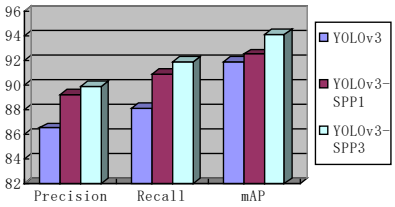

Fig. 4. Evaluation indexes of YOLOv3,YOLOv3-SPP1 and YOLOv3-SPP3 (Table 2 Input size is 864).

According to the comparison of evaluation indexes of YOLOv3, YOLOv3-SPP1 and YOLOv3-SPP3 in Figure 4, it can be seen that the value range of evaluation indexes rises with the increase of the number of spatial pyramids added to YOLOv3 detection head, which proves that adding spatial pyramid module to the detection head of YOLOv3 to enhance the extraction of data features has an impact. So YOLOv3-SPP1 and YOLOv3-SPP3 are the pruning models.

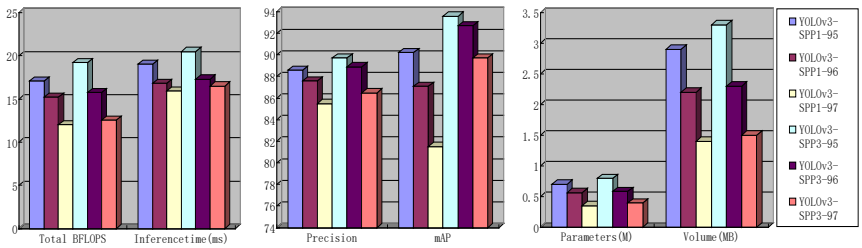

Fig. 5. Evaluation indexes of different pruning rates of YOLOv3-SPP1 and YOLOv3-SPP3 (Table 2, Input size is 864).

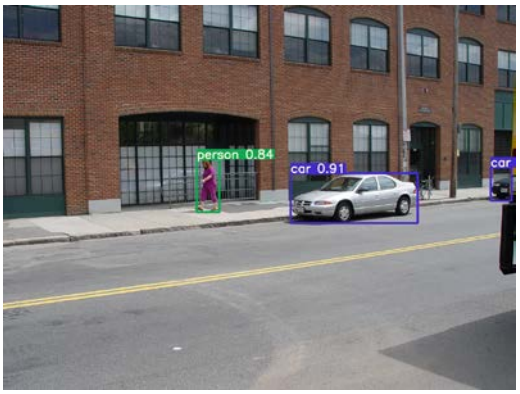

Fig. 6. YOLOv3 detection effect.

During the experiment,YOLOv3-SPP1 and YOLOv3-SPP3 were pruned in different proportions. It can be clearly seen from Figure 5 that the best effect was achieved when the pruning rate reached 96%, and the performance index of the algorithm model showed a trend of significant decline when the pruning rate increased to 97%.Based on the comparative analysis of the evaluation effect indicators of YOLOv3-SPP1-95, YOLOv3-SPP1-96,YOLOv3-SPP1-97,YOLOv3-SPP3-95, YOLOv3-SPP3-96 and YOLOv3-SPP3-97 in Figure 5, YOLOv3-SPP3-96 is selected as the optimal algorithm model YOLOv3-SPP3-tiny.

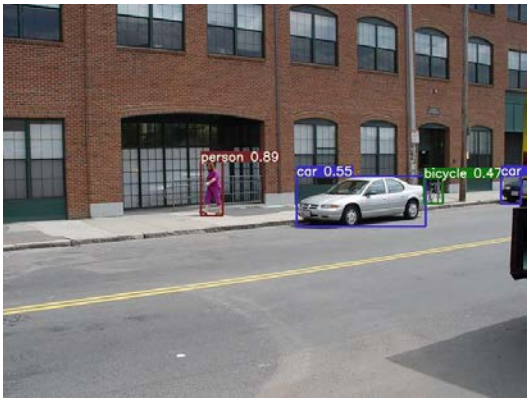

Fig. 7. Detection effect of YOLOv3-SPP3-tiny.

Analysis of model detection effect: Compared with YOLOv3-SPP3-tiny and YOLOv3 under the same training environment, the same network settings improve Precision by 2.4, Recall by 0.3, mAP by 0.8, Total BFLOPS by 269.91, Inference time by 70.90ms, Parameters by 59.94M, and Volume by 239.1MB Comparing the visual detection of YOLOv3 and YOLOv3-SPP3-tiny, as shown in fig.6 and fig.7, it can be found that there is no significant difference in detection effect. However, YOLOv3-SPP3-tiny reduced the time of reasoning each frame of pictures by 70.90ms, which greatly reduced the time of target detection and reserved more detection processing time for the detected target.

5 Conclusion

This experiment improves a visual detection algorithm YOLOv3-SPP3-tiny, which is an optimized target detection algorithm based on YOLOv3.In order to further improve the detection performance of YOLOv3, a spatial pyramid network was added at the end of the YOLOv3 network to improve the richness of deep features. Target detection algorithms related to deep learning take up large space volume and operation consumption. Therefore, in order to reduce the consumption of space and operation resources, a layer and channel fusion pruning method was developed to optimize the YOLOv3 algorithm model. In addition, in order to automatically remove the unimportant parts in the pruning process, L1 sparse regularization is first applied to the network, and the unimportant parts are removed according to the scaling factor of debugging. Based on the above optimization strategy, the YOLOv3-SPP3-tiny model is obtained. Compared with YOLOv3, the real-time performance and detection effect of YOLOv3-SPP3-tiny target detection are greatly improved, and the space volume of the YOLOv3-SPP3-tiny optimized model is reduced by 99.05% compared with YOLOv3.Therefore, YOLOv3-SPP3-tiny is more effective than YOLOv3 for driverless vehicles in complex urban traffic environments.

This work was supported in part by the Tianjin University Discipline Leading Talent Training Program of China (No. SSW181030105) and the Tianjin Artificial Intelligence Project of China (No. 18ZXAQSF00090).

References

1. H. Lu, Y. Li, T. Uemura, H. Kim, S. Serikama.. Low illumination underwater lightfield images reconstruction using deep convolutional neural networks, Future Gener. Comput. Syst. 82 (2018) 142-148.

2. H. Lu, Y. Li, M. Chen, H. Kim, S. Serikawa.. Brain intelligence: go beyond artificial intelligence. Mobile Networks and Applications, (2017), 23(7553), 368-375.
3. H. Lu, B. Li, J. Zhu, Y. Li, Serikawa.. Wound intensity correction and segmentation with convolutional neural networks, Concurr. Comput.: Pract. Exper. 29 (6) (2017).
4. P. Li, D. Wang, L. Wang, H. Lu.. Deep visual tracking: review and experimental comparison, Pattern Recogn. 76 (2018) 323-338.
5. C. Sun, D. Wang, H. Lu, M.H. Yang.. Learning spatial-aware regressions for visual tracking, 2018.
6. C. Sun, D. Wang, H. Lu, M.H. Yang.. Correlation tracking via joint discrimination and reliability learning, 2018.
7. J. Redmon, S. Divvala, R. Girshick, A. Farhadi.. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
8. J. Redmon, A. Farhadi..: Yolo9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 6517-6525 (2017)
9. J. Redmon, A. Farhadi..: "YOLOv3: an incremental improvement (2018)." arXiv preprint at https://arxiv.org/abs/1804.02767(1804)
10. S. Han, J. Pool, J. Tran, W.J. Dally.. Learning both weights and connections for efficient neural network. In NIPS, pages 1135-1143, 2015.
11. H. Li, A. Kadav, I. Durdanovic, H. Samet, H.P. Graf.. Pruning filters for efficient convnets. arXiv preprint arXiv:1608.08710, 2016.
12. Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, C. Zhang.. "Learning efficient convolutional networks through network slimming." Proceedings of the IEEE International Conference on Computer Vision (2017). https://doi.org/10.1109/ICCV.2017.298
13. K. He, X. Zhang, S. Ren, J. Sun.. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(9), 1904-1916 (2014)
14. Scardapane, Simone, Hussain, Amir, Comminiello, Danilo, Uncini, Aurelio.. Group sparse regularization for deep neural networks. Neurocomputing, 241, 81-89 (2017) arXiv preprint at https://arxiv.org/abs/1607.00485
15. Glenn Jocher, guigarfr, perry0418, Ttayu, Josh Veitch-Michaelis, Gabriel Bianconi, IlyaOvodov. (2019). ultralytics/YOLOv3:Rectangular Inference, Conv2d + Batchnorm2d Layer Fusion (Versionv6). Zenodo. http://doi.org/10.5281/zenodo.2672652
16. S. Ioffe, C. Szegedy.. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.