Vehicle tracking algorithm based on deep learning

Xiao Feng1*, Yan Piao1 and Sitong Sun1
1 School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, Jilin, China

*Corresponding author’s e-mail: 2019100552@mails.cust.edu.cn

Abstract. Vehicle detection is one of the most important detection targets in target detection. Overcoming this problem is of great significance in the field of traffic detection and automatic driving. In this paper, based on yolov4, firstly, the slice module is added to the input, and the slice module slices the image into the backbone to get the sampling feature map without information loss. Then, the SPP module is added to the neck of the neck module, and then the Kitti dataset is trained, detected and tested. Experiments show that adding slice module in the input circuit and spp module in the neck can effectively improve the accuracy of vehicle detection Map@0.5 3.4%.

1. Introduction
With the development of the city, the density of urban traffic flow and people flow has become an important reason for urban road traffic congestion. According to the traffic administration bureau of the Ministry of Public Security, by June 2020, the number of automobiles in China had reached 270 million, accounting for 75 percent of the total number of motor vehicles. A total of 12 cities had more than 3 million cars, with Beijing and Chengdu topping the list with more than 6 million and 5 million respectively. However, with the popularization of cars, there are more and more safety problems on the road, and traffic accidents are more and more frequent[1]. Drunk driving, fatigue driving, weather, road conditions and so on are the most common causes of traffic accidents, and automatic driving can effectively solve these problems. The advantage of mature automatic driving technology is not only to improve the safety of vehicle driving, but also to liberate the driver. As an important part of automatic driving, pedestrian detection can effectively detect the pedestrian near the vehicle and its position information[2]. Combined with other related technologies, it can predict the moving direction of the pedestrian and make a correct response in time.

The current target detection technology is mainly separated into two classes. One is the two-stage detection algorithm based on region generation represented by FASTR-CNN and FASTERR-CNN. The second is the regression based one-stage detection algorithm represented by YOLO and SSD. The two-stage detection algorithm usually has high detection accuracy, but the detection speed is slow, while the one-stage detection algorithm enhances the detection speed at the expense of certain detection accuracy[3].

At present, YOLO(You only Look once) has many applications in various fields of engineering. In terms of industry, Zhou Quan min and Du Yu jie et al. used YOLOv2, YOLOV3 and YOLOV4 to fine-tuning, training, verification and analysis the network model on the basis of Darknet framework for the detection of photovoltaic modules' defects (spot spot, multi-spot, strip spot and no-load). In the field of medicine, Wang Bo and Feng Xupeng et al. proposed a deep convolutional neural network with close connection between multiple scales based on YOLO algorithm and combining the ideas of Darknet-53 network and DenseNet network. It effectively improves the recognition accuracy and speed of pulmonary
nODULES IN LUNG CT IMAGES, AND PROVIDES THE CONDITIONS FOR REAL-TIME DETECTION OF PULMONARY NODULES IN LUNG CT IMAGES[4]. IN THE FIELD OF AVIATION AND AEROSPACE, PAN WEIJUN AND LIU HAOCHEN ET AL. PROPOSED A CONVOLUTIONAL NEURAL NETWORK BASED ON YOLO FRAMEWORK TO IDENTIFY AND DETECT AIRCRAFT IN SEA IMAGES CAPTURED BY UNMANNED AERIAL VEHICLES, AND TESTED THE PERFORMANCE OF THE LATEST DETECTION AND RECOGNITION METHODS UNDER SPECIAL CIRCUMSTANCES. THUS, YOLO SERIES DETECTION ALGORITHMS HAVE BEEN APPLIED TO ALL WALKS OF LIFE, AND ACHIEVED GOOD DETECTION RESULTS, WITH CERTAIN PRACTICAL APPLICATION VALUE.

THE SINGLE-STAGE DETECTION ALGORITHM YOLOV4 HAS BECOME THE MAINSTREAM DETECTION ALGORITHM IN ENGINEERING APPLICATIONS BECAUSE OF ITS GOOD DETECTION EXACTNESS AND SPEED. IN THIS PAPER, BASED ON THE KITTI DATA SET, THE YOLO_V4 ALGORITHM IS USED TO CONDUCT FEATURE TRAINING AND LEARNING ON THE DATA SET, AND FURTHER ADJUST THE PARAMETERS OF THE NETWORK MODEL, AND FINALLY OBTAIN THE VEHICLE AND PEDESTRIAN DETECTION MODEL IN THIS PAPER.

2. INTRODUCTION OF YOLOV4 ALGORITHM

YOLOV4 is a single-stage target detection algorithm with strong real-time performance. The algorithm consists of three parts: Backbone network for feature extraction, Neck Neck for feature fusion, and detection Head for classification and regression operation. Compared with the classical YOLOV3 target detection algorithm, YOLOV4 algorithm integrates the excellent algorithm model ideas and training skills in deep neural networks in recent years, based on the algorithm structure of "Darknet53+FPN+YOLO-HEAD" of YOLOV3. On the basis of Darknet, the model backbone network fuses the idea of CSPnet algorithm to form CSP Darknet, which achieves the effect of reducing network computation and ensuring the accuracy of network[5]. In the neck, the feature pyramid network (FPN) adopted by YOLOV3 was replaced by the Path Aggregation Network (PANET) with Spatial Pyramid Pooling (SPP), which transferred the deep layer features from the backbone network to the shallow layer, and improved the problem of the shallow layer feature loss caused by the transfer of shallow layer features from the shallow layer to the deep layer[6]. The detection Head continued the YOLO-head in YOLOV3, as shown in Table 1, and finally formed the model structure of "CSP Darknet+Pan-SPP+YOLO-head".

2.1. ALGORITHM DETECTION PROCESS

As shown in Figure 1, Yolov4 adjusts the size of the input image to 608×608 at the input end and inputs it into the network for training and detection. With CBM convolution layer and the residual Resunit stack of CSP module of backbone, in deepening the network layer get richer semantic information characteristics of figure on the basis of effective prevent the gradient disappear or explosion problem, and in the backbone network by step length is 2, 5 times for 3 layers of convolution kernel size under sampling to realize the dimension reduction of characteristic figure; Two times of upsampling were carried out in the network neck, and the PAN+SPP model structure was used to achieve the fusion of shallow layer features and high-level semantic features as well as the fusion of multi-scale receptive fields, which made full use of the detail features of the shallow layer network and improved the problem of small target feature loss. Detection by using the regression + classification head, the input image is divided into 76 x 76, 38 * 38, 19 lines of three different size of grid map, respectively to achieve the targets for small, medium and large target detection, compared to the two stage detection algorithm, YOLOv4 while effectively improve detection precision, save a large amount of computing resources and training time cost, improve the detection speed[7].
Fig 1. YOLOv4 network structure diagram

YOLOv4 network structure consists of the following five basic components, as shown in Fig 2:
(a) CBM module structure: the smallest component in yolov4 network structure, which consists of Conv, Bn and Mish activation functions.
(b) CBL module structure: composed of Conv, Bn and Leaky_relu.
(c) Result module structure: a deeper network structure that can be constructed by using the residual structure in Resnet network structure for reference.
(d) CSP module structure: referring to CSPNet network structure, it is composed of convolution layer and x Res unit modules Concat.
(e) SPP module structure: multi-scale fusion is carried out by adopting the maximum pooling mode of $1\times1$, $5\times5$, $9\times9$ and $13\times13$. 
2.2. Improved yolov4 algorithm

2.2.1. Add a Slice module to the input

The study of Kaiminghe et al. showed that the input of the full connection layer of the convolutional neural network must be a fixed eigenvector, and the direct stretching of the image would lead to the loss of image information, thus affecting the accuracy of recognition. SILCE module in the image into the Backbone for slicing operation, of slice images for operation, the operation is every pixel in an image China get a value, similar to the adjacent sampling, under the picture so that it can be divided into 4 pictures, 4 pictures complement each other, which will help in the process of the following sampling information loss, makes the image of W, H information into the channel space, input channel into four, at the age of information integration of the acquisition of four images in RGB three-channel pattern into twelve channels, finally will receive pictures again for convolution operation. The result is a double subsampled feature map with no loss of information. As shown in the figure (3) below, the original 608×608×3 image is input into the slice structure and is firstly transformed into a 304×304×12 feature image through slice operation. After a convolution operation of 32 convolution kernels, it is finally transformed into a 304×304×32 feature image.

Fig.2 The five components of YOLOv4 network structure
2.2.2. YOLOv4 algorithm with SPP module added

Inspired by the SPPNET network, the YOLOV4 model adds an SPP module after the backbone network, as shown in the figure.

As shown in Fig. 2, after the input feature map passes through a convolutional layer, it is processed by 5×5, 9×9 and 13×13 cores of different sizes for maximum pooling, and then channels are concatenated for the obtained feature map. The number of output channels is 4 times of the original number of channels, and the size of the feature map remains unchanged.

The MAXPool layer expands the receptive field while maintaining the translation invariance of the feature map, while the SPP module uses the MAXPool layer of different kernel sizes to obtain the receptive field information in the local region of the feature map and near the global receptive field information, and performs feature fusion. This fusion of receptive fields at different scales can effectively enrich the expression ability of the feature map, enhance the acceptance range of the output features of the backbone network, isolate important contextual information, and effectively improve the detection performance of the model.

In YOLOV4, the SPP module is located behind the backbone network with a feature graph size of 19×19, and the feature information of the SPP module is directly fed into the YOLO detection head with a feature graph size of 19×19 for result prediction. Based on the above basis, this paper respectively on YOLOv4 network for the first time after sampling and sampling on the second, the network layer to join the 123th floor and 140 SPP module, to increase the input to 38 * 38 and 76 * 76 size YOLO expression ability of detecting head figure characteristics of information, achieve better detection effect, this article will join SPP YOLOv4 network called YOLOv4 - 3SPP module of improvement, as shown in Figure 4.
3. Experiment and result analysis

3.1. Experimental platform

The experiment is based on anaconda3 platform and is programmed by python3.6 language. All experiments are completed on the laboratory server terminal configured with Intel Xeon e52678v3 CPU and NVIDIA 1080ti GPU, the operating system is Windows Server 2016 64 bit and memory is 64GB. The maximum learning efficiency set in this experiment is 0.001, attenuation coefficient is 0.005, and the size of the picture set in the experiment is 608 × 608, batch Set the size to 6 and the training epochs to 30000.

3.2. Data sets

The data set uses the public Kitti target detection benchmark data set and Kitti target tracking benchmark data set. The target detection data set is mainly for the detection of vehicles and pedestrians, and its training set and test set have 7481 and 7518 pictures respectively, while the target tracking data set is mainly for the tracking of vehicles and pedestrians, and its training set is mainly for the detection of vehicles and pedestrians There are 21 and 29 video sequences in the training set and the test set respectively. In this experiment, the original eight different categories labeled in the Kitti target detection benchmark data set are merged into two categories. Specifically, car, van and truck are merged into car categories, while pedestrian and person are merged into car categories. At the same time, the original annotated pedestrian and person classes of Kitti target tracking benchmark dataset are merged into pedestrian, and van and car are merged into car classes, and only car, pedestrian and cyclist classes are retained 1 : 1 is divided into training set, verification set and test set for training and evaluation of target detection network.

3.3. Experimental result

In this experiment, four performance indicators, precision (P), recall (R), F1 score, map (mean average precision, and average AP value), are used to evaluate the network performance. Map is used to evaluate the multi label image classification task, and it is an important indicator to measure the overall detection accuracy of the model in multi category target detection. The precision P, recall R and map are expressed as:

\[
P = \frac{X_{TP}}{X_{TP} + X_{FP}} \quad (1)
\]

\[
R = \frac{X_{TP}}{X_{TP} + X_{FN}} \quad (2)
\]

\[
mAP = \frac{\sum_{i=1}^{c} AP_i}{c} \quad (3)
\]
Where: $X_{TP}$ denotes the number of correctly detected targets; $X_{FP}$ denotes the number of targets detected by error; $X_{FN}$ denotes the number of targets that have not been detected; $C$ denotes the number of categories, and AP is the average precision of a single target category, where $AP = \int_0^1 P(R)dR$.

4. Conclusion
This paper uses the improved yolov4 algorithm to detect and track vehicles. As shown in Table 1, the average accuracy of vehicle recognition is 88.5%, and the real-time FPS reaches 30 FPS, which meets the requirements of real-time vehicle detection. By improving the network model of yolov4, the detection accuracy is significantly improved compared with the original algorithm, as shown in figures.

| Algorithm     | P   | R   | FPS | Map@0.5 |
|---------------|-----|-----|-----|---------|
| YOLOv4        | 85.5% | 54.4% | 20  | 85.2%   |
| YOLOV4-3SPP   | 88.6% | 65.4% | 30  | 88.5%   |

![Comparison of yolov4 and improved algorithm](image)

Fig 5. Comparison of some experimental results

References
[1] Zhao J. Road traffic safety evaluation index system based on complex system-entropy theory[C]// International Conference on Electric Technology & Civil Engineering. IEEE, 2011.
[2] Wang E, Y Li, Sun A, et al. Road detection based on illuminant invariance and quadratic estimation[J]. Optik - International Journal for Light and Electron Optics, 2019, 185:672-684.
[3] Shin H, Fan X. Road vanishing point detection using weber adaptive local filter and salient-block-wise weighted soft voting[J]. Iet Computer Vision, 2016, 10(6):503-512.
[4] Zw A, Jx B, Ps A, et al. Improved lung nodule diagnosis accuracy using lung CT images with uncertain class[J]. Computer Methods and Programs in Biomedicine, 2018, 162:197-209.
[5] Shao J, Qu C, Li J, et al. A Lightweight Convolutional Neural Network Based on Visual Attention for SAR Image Target Classification[J]. Sensors, 2018, 18(9):3039.
[6] Wang-wang, ZHOU, Jie-yong, et al. Recognition of Combat Intention with Insufficient Expert Knowledge[C]// 2018.
[7] Lopez-Martin M, Carro B, Sanchez-Esguevillas A, et al. Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things[J]. IEEE Access, 2017, PP(99):1-1.