Human-vehicle detection based on YOLOv5

Zhihui Chen[1], Xiaoning Yan[2], Shuangwu Zheng[2], Xiaoyan Chen*[1]

[1]Tianjin University of Science and Technology, China;

[2]Shenzhen Softsz Co. Ltd., China;

E-mail: 1594838831@qq.com
www.tust.edu.cn

Abstract

With the continuous improvement of social development level, traffic has become complicated. Therefore, the detection of people and vehicles becomes important. There are many application scenarios for human-vehicle detection, such as autonomous driving and transportation. This paper mainly introduces the research status of human-vehicle detection, analyzes the advantages and disadvantages of various current target detection algorithms, and focuses on YOLOv5 algorithm. Because the YOLOv5 model is much smaller than YOLOv4, and YOLOv5 also has strong detection ability. Finally, YOLOv5 is used to carry out human-vehicle detection experiments. The results the detection accuracy is improved slightly.

Keywords:human, vehicle, detection, YOLOv5

1. Introduction

At present, target detection has not only received a lot of research in academia, but also has been widely used in real life, such as video fire detection, unmanned driving, security monitoring, and drone scene analysis. Target detection algorithms are mainly divided into three categories: traditional target detection algorithms, classifier-based detection algorithms and regression-based detection algorithms.

In terms of pedestrian and vehicle detection. Initially, the pedestrian detection algorithm mainly used independent single features for feature selection in videos and pictures, and it was mainly aimed at pedestrians in social streets[1]. The most commonly used feature selection in China is the histogram of directional gradients (HOG). Its core point is that some appearances and features of the detected object can be better represented by the distribution of gradient and edge direction. Traditional foreign target detection methods mainly explore target classification, such as how to distinguish pedestrians from other objects in the street and how to use classifiers reasonably. In vehicle detection, the more commonly used method is for researchers to establish information and non-redundant derived values from the initial measurement data. The advantage of this method is to promote learning and improve the generalization ability of the model, and it can also bring better interpretability.

2. Target Detection Algorithms

Target detection algorithms can be divided into traditional target detection algorithms, classifier-based detection algorithms and regression-based detection algorithms.

2.1. *Traditional target detection algorithm*

The SIFT algorithm can extract its invariant characteristics from the features when the image proportion and rotation are unchanged.

AdaBoost is an iterative algorithm that adds a new weak classifier in each round until it reaches a predetermined and sufficiently small error rate.

2.2. *Detection algorithm based on classifier*

OverFeat is an early stage one-stage target detection method. This method mainly discusses multiscale and sliding window methods. It can be used in a feedforward network that includes convolution calculations and has a deep structure. This method can better locate the target by determining the boundary information of the target in advance.

R-CNN (Region with CNN features) algorithm is a type of recurrent neural network that takes sequence data as input, recursively in the evolution direction of the sequence, and all nodes (recurrent units) are connected in a chain.

Fast-RCNN algorithm is established on the basis of previous research. It uses a feedforward neural network that includes convolution calculations and has a deep structure, which can be used to effectively classify candidate targets. Compared with previous research, Fast-RCNN has many innovations, which improves the speed of training and testing while also improving the accuracy of detection.

2.3. *Regression-based detection algorithm*

YOLO (You Only Look Once) is an algorithm that can solve bounding boxes with accurate predictions while using convolutional sliding windows[2]. The core idea of YOLO is to transform target detection into a regression problem, using the entire image as the input of the network, and only going through a neural network to get the location of the bounding box and its category. The YOLO algorithm can realize real-time detection.

Compared with the YOLO algorithm, the SSD (Single Shot MultiBox Detector) algorithm directly uses CNN for direct detection. The algorithm uses convolutional feature maps of different scales for detection. Large-scale feature maps can be used to detect small things, and small-scale feature maps can be used to detect large objects, so that objects of different scales can be detected.

3. YOLOv5 and Improvements

3.1. *YOLOv5*

Compared with YOLOv3, YOLOv5 has innovated in four parts of the network structure.

3.1.1. Input

(1)Mosaic data enhancement
This method randomly zooms, cuts, and arranges four pictures randomly, turns them into a new picture, and then puts the new picture into the network for learning.
(2)Adaptive anchor frame
In the early stage of training, a predefined frame will determine the location of the target at a possible location. As the training progresses, the real frame will gradually shift based on the preset frame for construction. Calculate the best anchor frame value in different training sets adaptively.

3.1.2 Backbone

The focus layer is very similar to adjacent downsampling. Suppose there is a 4x4 picture that is concatenated with separated pixel values into four 2x2 pictures. This structure can avoid information loss, the number of channels has become 4 times the original, and the size is half of the original.

3.1.3 Neck

Inserting the Neck layer between the input layer and the

Backbone layer improves the ability to extract fusion features. Can better detect targets of different sizes.

3.1.4 Output

Bounding box is used in YOLOv5. Bounding box is to fine-tune the predicted box to make it close to the ground truth box.

3.2. *Improvement*

3.2.1 BN and RBN

Batch norm is batch normalization[3]. It is to complete the normalization by adding parameters during the training process to solve the problem of normalized learning features. When applying Batch Norm, it should be satisfied that the mutual characteristics of different instances completely obey the same distribution. However, the distribution characteristics of the test set instances and the training data cannot be consistent. Therefore, the inconsistency in the training and testing process will weaken the actual effect of the BN layer.

In order to solve the above problems, this article quotes the method of RBN(Representative Batch Norm). RBN mainly takes two main steps: centering calibration and scaling calibration.

(1)Centering Calibration

$$X_{cm(n,x,h,w)} = X_{（n,c,h,w）} + W_m \cdot K_m \qquad （1）$$

In Eq.(1), X is set as the input feature W_m is set as a learnable weight vector, and K_m is expressed as an example feature. . This article uses GMP(Global Max Pooling). The learnable variable Wm is set to (N, C, 1, 1). '·'stands for dot product operation, which mainly converts two features into the same shape and then performs dot product operation.

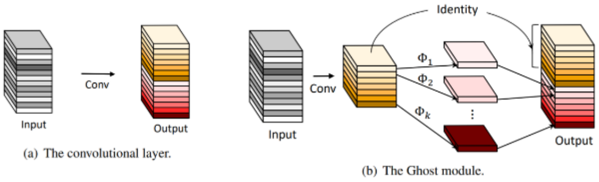
(2) Scaling Calibration

In the next step of Batch Norm, specifically, before the stretching adjustment, do the image scaling and alignment related operations.

In the experiment, the BN layer was replaced with RBN to test its training effect in the YOLOv5 network.

3.2.2 Ghost model network

The core idea of GhostNet is to use cheap operations to replace ordinary convolution operations to generate these redundant feature maps. The Ghost module divides the ordinary convolution into two parts. First, it performs an ordinary 1x1 convolution, which is a small amount of convolution. For example, a 32-channel convolution is normally used. Here, a 16-channel convolution is used. The effect of this 1x1 convolution Similar to feature integration, the feature enrichment of the input feature layer is generated. Then we perform deep separable convolution. This deep separable convolution is layer-by-layer convolution, which is the cheap operations we mentioned above[4]. It uses the feature enrichment obtained in the previous step to generate a Ghost feature map. Using the Ghost Net network can reduce the amount of model parameters and increase the execution speed of the model while ensuring a good detection effect.The Ghost model is shown in the Fig 1 below.



(a) The convolutional layer.

(b) The Ghost module.

Fig.1. Ghost model schematic

4. Experiment and Result Analysis

4.1. *Experimental equipment*

The experimental equipment is shown in Table 1.

Table 1. Device-related configuration

Name	Version and Model
Ubunt version:	20.04.3 LTS
CUDA version:	11.2
Graphics:	GeForce GTX 1080 Ti
Frame:	Torch

4.2. ***Data collection***

The preliminary data preparation mainly calibrates two categories, people and cars. It need to copy the .txt files with several preset classes to the relevant folder of the labelimg software. Through repeated screening and data supplementation, 30938 images of the data set and corresponding label files are provided, script codes are written, and the training set and the test set are allocated according to a 9:1 ratio. There are 27761 training sets and 3177 verification sets. Since the label file contains multiple categories(the experiment only recognizes people and cars, two categories), write script code to retain the specified label content (0-person, 1-car).

4.3. ***Experimental setup***

Set lr=0.01, batch size to 16, and epoch to 300. The remaining parameters are default values for experimentation. Use the mAP (mean Average Precision) as the evaluation index.

Next, select about 500 pictures as the detect data and calculate the average inference speed. The experimental results are shown in Table 2.

Table 2 .Model size comparison

Model	mAP_0.5	mAP_0.5:0.95	Time(s)
YOLOv5s	0.9072	0.63213	0.02813
YOLOv5m	0.9278	0.67583	0.02879

It is known that the depth and width of the YOLOv5m network is greater than that of YOLOv5s. Combined with Table 2, it can be seen that increasing the network depth and width will increase the mAP value, but it will also increase its reasoning time. Therefore, this experiment uses the faster YOLOv5s for training and comparison.

4.4. ***Experimental result***

The size of the convolution kernel restricts the amount of calculation of its model. Adjust the depth to separate the convolution kernel size, set g=1 and set the default kernel size to 5*5, respectively, adjust to 1*1 (point-by-point volume Product), 3*3 observe its model size and performance. The experimental results are shown in Table 3.

Table 3 .Network performance of different sizes of deep separable convolution kernels in Ghost mode

Convolution size	FLOPs	Time(H)	Parameters (M)	Model size	mAP@.5 mAP@.5:.95
1*1	14.5	14.7	6.283	27.8MB	0.900;0.626
3*3	14.5	16.6	6.287	27.8MB	0.901;0.629
5*5	14.6	19.4	6.295	30.8MB	0.905;0.632

As shown in Table 3, different convolution kernel sizes affect the amount of parameters and the size of the mAP. The larger the convolution kernel, the larger the mAP, and the amount of parameters will increase accordingly. The human-vehicle detection effect of the YOLOv5 network is shown in the Fig 2.

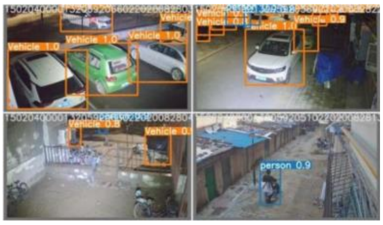


Fig.2. The result of recognition by model

5. Concluding

In terms of data, the universality can be improved by adding data types, for example, not only for learning and testing vehicles in a specific area, but also for learning and using vehicle data from all parts of the country or the world within the scope of the conditions.

In terms of algorithm selection, although the mainstream YOLOv5 network is relatively complete and has been widely used, some lightweight improvement ideas (pruning, distillation) can still provide better support for the use of this network on the mobile side, making the model size smaller. The complexity of calculation is reduced, the energy consumption of the mobile terminal is reduced, and the flexibility of deployment is improved[5]. A single picture containing a large number of people and cars can introduce the idea of clustering to group data points that are relatively close together. In addition, the transformer that shines in the field of NLP can now be applied to target detection. It provides a new idea for target detection to transform the detection problem into a set-based index problem, which is different from the traditional replacement of the backbone or the addition of a special FPN. If the attention mechanism can be properly combined with the YOLOv5 network, there may be better performance. This is also a research direction in the future.

References

1. Y. Hou, Y. Song, X. Hao et al., "Multispectral pedestrian detection based on deep convolutional neural networks," 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017, pp. 1-4, doi: 10.1109/ICSPCC.2017.824250
2. W. Li, "Infrared Image Pedestrian Detection via YOLO-V3," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 1052-1055, doi: 10.1109/IAEAC50856.2021.9390896.
3. W. Lan, J. Dang, Y. Wang et al., "Pedestrian Detection Based on YOLO Network Model," 2018 IEEE International Conference on Mechatronics and Automation (ICMA), 2018, pp. 1547-1551, doi: 10.1109/ICMA.2018.8484698.
4. Y. Huang, Y. Zhou, J. Lan, et al., "Ghost Feature Network for Super-Resolution," 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), 2020, pp. 1-3, doi: 10.1109/CSRSWTC50769.2020.9372549.
5. J. Tang, S. Liu, B. Zheng, et al., "Smoking Behavior Detection Based On Improved YOLOv5s Algorithm," 2021 9th International Symposium on Next Generation Electronics (ISNE), 2021, pp. 1-4, doi: 10.1109/ISNE48910.2021.9493637.

Authors Introduction

Mr. Zhihui Chen

He received his bachelor's degree from the school of electronic information and automation of Tianjin University of science and technology in 2021. He is acquiring for his master's degree at Tianjin University of science and technology.

Mr. Xiaoning Yan

He obtained a bachelor's degree from South China University of Technology. He obtained a master's degree in computer science from the University of Texas at Dallas from 2013 to 2015.

Mr. Shuangwu Zheng

He is a graduate student in control engineering of Tianjin University of Science and Technology. He is currently the deputy dean of Ansoft Algorithm Research Institute and an algorithm engineer of Ansoft Technology.

Prof. Xiaoyan Chen

She, professor of Tianjin University of Science and Technology, graduated from Tianjin University with PH.D (2009), worked as a Post-doctor at Tianjin University (2009.5-2015.5).