The Research on autopilot system based on lightweight YOLO-V3 target detection algorithm

Xiulong Gao, Dongyuan Ge and Zhifeng Chen
School of Mechanical and Transportation Engineering, Guangxi University of Science and Technology, Liuzhou, Guangxi, 545006, China
E-mail: 1029811078@qq.com

Abstract. Computer vision is one of the most important branches in the field of deep learning, which includes image classification, target detection and so on. In this paper, the pruning algorithm is combined with the method of reducing the network depth, and the trained network model is applied to the field of autopilot, which solves the problems of high power consumption, single function and poor real-time performance of the vehicle target detection system. In this paper, through a large number of theoretical research and experiments, the appropriate activation function and regularization method are selected. Through the visualization method of neural network convolution layer, the working principle of the network is expounded in detail. Through theoretical analysis, the appropriate anchor box, is selected and the non-maximum suppression algorithm is introduced. Tensorboard data visualization system is used to monitor the distribution and law of network parameters. The accuracy of detection according to the experimental target is 98.7%.

1. Introduction
In recent years, with the continuous research and development of autopilot vehicles at home and abroad, computer vision technology and speech recognition technology have begun to be widely used in autopilot vehicles. The task of target detection in computer vision is to find out the targets of interest in each frame of a single image or video, and to determine their central coordinates, height, width and category relative to the image[1]. In view of the different colours, shapes, posture, texture and exposure of the detected objects, and the interference of occlusion and other complex factors, target detection has been one of the most challenging topics in the computer field since its emergence.
In view of the above problems, this paper puts forward the corresponding improvement to the YOLO-V3[2-3] algorithm, not only the real-time performance of the network processing information has been improved, but also the accuracy of object recognition has been improved, in order to meet the requirements of the autopilot system.

2. The architecture design of lightweight YOLO-V3 network
YOLO-V3 algorithm uses 53 layers network for feature extraction, the main disadvantage is that the network structure is complex and the weight parameters are redundant, which cannot meet the high real-time requirements of autopilot system. In this paper, a method of extracting feature information using 16 convolution layers is proposed. In CNN, the network uses the convolution kernel weight matrix of each layer to extract the features of the input training image, so in CNN, the number of convolution kernels and the size of convolution kernels are the main parameters of convolution layer.
Because the pixel matrix dimension of the original image input into the network is low, a small amount of convolution kernel can extract the input image information effectively and accurately in shallow network. Figure 1 is the working principle diagram of convolution kernel.

![Convolution Filter Diagram](image)

**Figure 1.** The working principle of convolution kernel.

In view of the above theoretical analysis, 64 convolution cores are used for layer 1.2 of the network. According to the high-dimensional data processed by the former shallow network, the layer 3 and 4 networks use more than 128 convolution kernels. Through the increasing number of convolution cores, the information of the original image is decomposed continuously. Because the shallow network continues to input high-dimensional data into the deep network, the number of convolution kernels needed in the network increases layer by layer. A large number of experiments and theoretical analysis show that the convolution kernel size is 3, when the size is too large, the extracted feature information contains too many feature elements. when the convolution kernel size is too small, the extracted feature information is too fine to be used as a classification reference, and then the performance of the network is degraded[4].

### 3. The selection and optimization of main parameters of CNN

#### 3.1. The selection and Optimization of Activation function

Relu activation function: when the training data distribution is greater than 0, the gradient does not disappear and is easy to calculate, and when the training data distribution is less than 0, the neurons are not activated, which is helpful to the establishment of network sparsity. Relu functions (1):

\[
f(x) = \max(x, 0)
\]

In order to solve the problem of ReLU, the Leaky ReLU activation function is proposed. The main idea is that when \( x < 0 \) is multiplied by a relatively small coefficient, the function expression is:

\[
f(x) = \max(\alpha x, x)
\]

The \( \alpha \) in Leaky ReLU is a small constant, which is assigned according to the experimental experience. Compared with ReLU, the parameter is introduced into the negative half axis of \( x \), so that the reciprocal of the function is not 0 in the region of \( x < 0 \), so that all the distributed data can be fully utilized. The experimental results in tensorflow-playground show that Leaky ReLU performs best, so the activation function selects Leaky ReLU. The following Figure 2 shows how the activation function works.
3.2. The selection of gradient drop algorithm

SGD algorithm is the main optimization algorithm in the field of deep learning because of its fast computing speed and the less need of the hardware resources. The formula is as follows.

\[ g(\phi) = \sum_{j=1}^{n} \phi_{j}x_{j} \]  

(3)

\[ h(\phi) = \frac{1}{2m} \sum_{i=1}^{m} \left( y_{i} - g_{\phi}(x_{i}) \right)^{2} \]  

(4)

\[ \phi := \phi - \eta \nabla h(\phi) \]  

(5)

In the formula, \( \phi \) represents the weight in the network, \( \nabla \phi \) represents the gradient value, \( h(\phi) \) stands for the loss function, \( g(\phi) \) represents the objective function, \( y_{i} \) represents the sample value of the first sample, \( m \) represents the number of iterations, \( \eta \) represents the learning rate, and \( j \) represents the total number of parameters in the network. Through the comparative experiment found that the learning rate is 0.02, the decrease rate of the loss value of the function is stable, and the fluctuation range of the loss function becomes smaller, so it is easier to reach the optimal solution.

4. The parameter setting of the target detection box

4.1. The parameter design of Anchor box

In view of the fact that the system is applied to an automatic driving system, the accuracy requirement of the positioning and classification of the trained CNN network model is high, an anchor box is introduced in order to avoid the problem that a plurality of objects cannot be detected in the same grid.
The benefit of creating an anchor box[5-6] is that the anchor box enables your learning algorithms to be more targeted, especially as the data set in this article has some narrow, high-profile objects, such as pedestrians and bicycles, and car-wide objects. According to the characteristics of the training data in this paper, six types of anchor box (18,17), (42,27), (49,40), (86,39), (119,80), (182,155) with different aspect ratios are set. Figure 3 is a schematic diagram of anchor box.

4.2. Non-maximum suppression
When CNN network model detect the target in the process of detection, the same target will be detected many times. In order to solve this problem, the non-maximum suppression prediction algorithm is introduced, and the redundant detection box is removed by setting the threshold and using the comparative IOU method. The specific method is to arrange the results of CNN network prediction from high to low. The other prediction boxes with low score are compared with the prediction box with the highest score, and the prediction box with IOU value greater than or equal to the threshold is removed, then the highest score is saved and the output is the most, and then the above algorithm is repeated in the remaining prediction box. It is more appropriate to set the threshold to 0.6 according to a large number of experiments and theoretical analysis. The working principle is shown in Figure 4.

![Figure 4. The working principle of non-maximum suppression algorithm.](image)

5. Experimental analysis
The network framework of this paper uses keras as the front end, tensorflow as the back end, computer CPU is i9 ≤ 9900K, 8 cores and 16 threads, the main frequency is 3.6 GHz, GPU: NVIDIA GeForce GTX1080TI memory 11 GB, memory 32GB, operating system ubuntu14.02. The prepared training data are enhanced and fed into CNN. After 200 iterations, the accuracy of the model on the verification dataset is 98.7%, and the decrease of loss function is as shown in figure 1. Through the analysis of Tensorboard, it can be seen that the weight distribution of the network is gradually dispersed by the density of initialization, which proves that the selection of network parameters is reasonable. The experimental results are shown in Figure 5,6,7.

![Figure 5. Accuracy and loss function change curve.](image)
Figure 6. The network weight parameter distribution change.

Figure 7. The experimental results of target detection.

6. Conclusions
From the theoretical analysis, it can be seen that the accuracy and operation speed of YOLO target detection network are inevitably related to the complexity of network structure and network parameters. On the basis of ensuring the accuracy, this paper reduces the network structure to reduce
the network complexity, and the processing information speed increases, which is better suitable for autopilot system.

Acknowledgments

The work described in this paper is partially supported by National Natural Science Foundation of china under grant Nos.51765007, the Guangxi Provincial Natural Science Foundation of China under grant No.2016GXNSFAA380111. the authors would like to thank the reviewers for their constructive comments and substantial help that improved the presentation of the paper.

References

[1] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. (2016) The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779-788.
[2] Redmon, J., Farhadi, A. YOLO9000: Better, Faster, Stronger. (2017) The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263-7271.
[3] Redmon, J., Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767, 1804.
[4] Simonyan, K., Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[5] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.
[6] Yu-Wei, C., Vijayanarasimhan, S., Seybold, S., Ross, D.A., Jia, D., Sukthankar, R. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. (2018) The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1130-1139.