Research on the Application of YOLOV4 Target Detection Network in Traffic Scenarios by Machine Vision Technology

Xingxing Li*, Panpan Yin, Chao Duan and Shuyue Zhang
School of Engineering, Guangzhou College of Technology and Business, FoShan, China

*Corresponding author: wslxx@jxstnu.edu.cn

Abstract. Deep learning target detection algorithm is maturing with the development of machine vision technology. This article applies YOLOV4 target detection algorithm to the industrial traffic scene to detect and recognize cars, trucks, buses, pedestrians, two-wheelers, and three-wheelers in traffic scenes. Six types of common goals. The algorithm in this paper cuts the network model structure, uses 12 anchors of different sizes to adapt to different size targets in the image, the accuracy of the trained detection model can reach 94.2%, and the tensorrt acceleration technology is used to deploy the network in NVIDIA Jetson AGX Xavier embedded edge computing device, and the algorithm inference speed can reach real-time 10ms per frame.

Keywords: Target detection algorithm, machine vision technology, YOLOV4.

1. Introduction
With the breakthrough of deep learning technology, artificial intelligence has begun to develop rapidly around the world at an unprecedented speed, and target detection algorithms based on deep learning technology have been widely used in industrial projects. For example, in traffic scenes, especially in complex road conditions, it is very time-consuming and labor-intensive to analyze traffic conditions only by staring at the real-time monitoring video loaded on the road conditions. With the help of artificial intelligence algorithms, real-time detection can be achieved. Analyzing information such as pedestrians and vehicles in a traffic scene can greatly improve the efficiency of human eyes searching for information, and it is not easy to miss effective road condition information.

In a complex traffic scene, there will be a large number of pedestrians and different types of vehicle information. These objects tend to move in real time, with varying degrees of occlusion, and the influence of interferents of traffic scene also have bright and dark lights at different times according to the weather. If you want to capture real-time road condition information in a traffic scene, you first need to be able to detect objects in the traffic environment in real time under complex conditions such as various weather and time periods. Therefore, you need to use a target detection algorithm that balances speed and detection accuracy. Landing deployment.

The development of the detection network model in 2016 has basically formed two network design modes, one-stage and two-stage. The common point of the two is that they both adopt the anchor-based design structure to achieve the effect of traversing the input feature map. But the phenomenon reflected is that the accuracy of the two-stage network is higher, and the speed of the one-stage network is faster.
MS-CNN [1] designs target detectors of different scales for different output layers, completes the detection problem at multiple scales, and uses the up-sampling of features to replace the up-sampling of the input image. Design a deconvolution layer to increase the resolution of the feature map so that small targets can still be detected. The de-convolutional layer of the feature map is used here to replace the up-sampling of the input image, which can greatly reduce the memory usage and increase the speed. DSSD [2] adds a deconvolution structure to the network, and improves the expressive ability of shallow features by using the ResNet structure in the backbone. YOLOV2 [3] improved YOLOV1 [4], drawing on network improvement techniques such as anchors and multi-feature layer fusion detection.

Under the premise of ensuring the detection speed, it improved the detection accuracy of the YOLO series. On the basis of YOLOV1, use anchors to strengthen the grid, increase the resolution of the input, replace dropout with BN, constrain the change interval of the center point of the anchor, and adopt a new backbone. The author of the FPN paper attempts to further enhance the detection accuracy of the network by enhancing the features of the output of the CNN backbone network. The CNN target detection network began to try to use multi-layer feature fusion to detect large targets plus small targets. The paper mainly proposes a new layer-jumping feature fusion and a method for classification. The proposal of the FPN network has also become a common structure of the backbone network for subsequent detection. The author of Mask-RCNN [5] tried to realize the detection task from the idea of using segmentation network, that is, to determine the category of each pixel, and then determine its minimum bounding rectangle through different examples to achieve the purpose of detection. YOLOV3 [6] continues to optimize and improve the YOLOV2 network, mainly using FPN and ResNet to improve the feature layer representation ability of the backbone network. Cascade R-CNN is also an optimization technique in the network training process. The author found that when training the detection network, it is necessary to set the hyper-parameter IOU threshold to determine whether the current positioning frame is a positive sample, but a single IOU threshold may not be suitable, so try A cascaded IOU threshold is used to assist training. In order to optimize the impact of a single IOU problem in RPN on the final detection accuracy, it is proposed to do a cascade of different IOU thresholds to improve the quality and ratio of positive and negative samples for calculating the final loss, thereby improving performance.

This paper cuts based on the YOLOV4[7] model, uses 12 anchors of different sizes to adapt to different sizes of targets in the image, and performs anchor clustering on the training data, and uses tensorrt to accelerate model quantification and deployment on the edge computing device, the real-time performance of the algorithm is guaranteed.

2. Principle

Compared with YOLOV3 [6], the network structure of YOLOV4 [7] has more CSP structure and PAN structure, as shown in Fig.1:
Among them, the full name of CSPNet is Cross Stage Parital Network, as shown in Fig.2. It mainly solves the problem of large amount of calculation in reasoning from the perspective of network structure design. CSPNet believes that the problem of excessive inference calculations is caused by the duplication of gradient information in network optimization. Therefore, the CSP module is used to first divide the feature map of the base layer into two parts, and then merge them through a cross-stage hierarchical structure, which reduces the amount of calculation while ensuring accuracy. Therefore, Yolov4 adopts the CSPDarknet53 network structure in the backbone network Backbone, which has three main advantages: 1. Enhance the learning ability of CNN, making it lightweight while maintaining accuracy; 2. Reduce computing bottlenecks; 3. Reduce memory costs.
The CSPNet structure splits the stack of the original residual block into two parts: the main part continues to stack the original residual block, and the other part is like a residual edge, which is directly connected after a small amount of processing. To the end. Therefore, it can be considered that there is a large residual edge in CSP.

2.1. Mosaic data enhancement
The Mosaic used in Yolov4 refers to the CutMix data enhancement method proposed at the end of 2019, but CutMix only uses two pictures for stitching, while Mosaic data enhancement uses 4 pictures, random scaling, random cropping, and random arrangement. Make splicing.

![Mosaic data enhancement](image)

**Figure 3.** Mosaic data enhancement.

In normal project training, the AP of small targets is generally much lower than that of medium and large targets. The Coco data set also contains a large number of small targets, but the more troublesome thing is that the distribution of small targets is not uniform. The use of Mosaic data to increase mainly has the following advantages: 1. Rich data set: random use of 4 pictures, random zoom, and then random Distributed splicing greatly enriches the detection data set, especially random scaling adds a lot of small targets, making the network more robust. 2. Reduce GPU: Some people may say that random scaling and ordinary data enhancement can also be done, but the author considers that many people may only have one GPU, so when Mosaic enhances training, it can directly calculate the data of 4 pictures, making Mini- The batch size does not need to be large, a GPU can achieve better results.

2.2. Mish activation function
YOLOV4 modified the activation function of DarknetConv2D from LeakyReLU to Mish. The Mish activation function formula and image are shown in Fig.4:

![Mish activation function](image)

**Figure 4.** Comparison of Mish and leakyReLU functions.
During the experimental test of the author of Yolov4, using the CSPDarknet53 network to do the image classification task on the ImageNet dataset, it was found that the accuracy of TOP-1 and TOP-5 using the Mish activation function were slightly higher than when they were not used. The Mish activation function formula is as follows:

\[ Mish = x \times \tanh(ln(1 + e^x)) \]  

(1)

2.3. Drop Block
The Drop block used in Yolov4 is actually similar to the Dropout function in common networks, and it is also a regularization method to alleviate overfitting. The traditional Dropout is very simple, it can be said in one sentence: Random deletion reduces the number of neurons and makes the network simpler. And Dropblock is similar to Dropout, as shown in Fig.5 below:

![Figure 5. Schematic diagram of Dropblock principle.](image)

The dropout method in the middle will randomly delete and discard some information, but the researchers of Dropblock believe that the convolutional layer is not sensitive to this kind of random discarding, because the convolutional layer is usually three layers together: convolution + activation + pooling layer, The pooling layer itself acts on adjacent units. And even if it is randomly discarded, the convolutional layer can still learn the same information from adjacent activation units. When the researchers of Dropblock compared and verified with Cutout, they found that there are several characteristics: Dropblock is better than Cutout; Cutout can only act on the input layer, while Dropblock applies Cutout to every feature map in the network; Dropblock Various combinations can be customized, and the probability of deletion can be modified at different stages of training. From the spatial and temporal levels, there are more refined improvements compared to Cutout.

3. Experiments
This experiment uses video images collected and organized by a traffic monitoring camera. The training images include images at different times such as early morning, morning, noon, afternoon, and night, as well as images under different weather conditions such as rainy, foggy, and sunny days. A total of 20642 training images. The detected targets are divided into 6 categories, including cars, buses, trucks, pedestrians, two-wheelers, and tricycles. After YOLOV4 training, the detection results are shown in Fig.6:
Fig. 6 shows the use of yolov4 to detect 6 types of targets in a traffic scene. The image in the upper left corner is the result of the night detection, and the rest are the detection conditions during the day. It can be seen whether it is a large target such as a bus or a truck. Small targets such as cars in the distance can be detected well, and the detection frame is relatively close to the target.

In this paper, the detection accuracy is quantitatively evaluated, tested on the 1987 test set, and the statistics of mAP indicators are shown in Fig. 7:

Fig. 7. Target detection accuracy mAP index.

The left image in Fig. 7 shows the label information of the image. The number of detection frames for each type of target is counted. It can be seen that the proportion of two-wheeled vehicles, three-wheeled vehicles, pedestrians, and buses is relatively small, and the sample is not balanced. The figure shows the AP value of each column of target detection. The category with the lowest AP value is pedestrians, reaching 85%, and the rest of the target AP values are all above 93%, indicating that the algorithm has a better performance on images in traffic scenes. Good detection effect. Finally, this article uses the tensorrt acceleration library to be deployed on the edge computing device, and the processing speed of the algorithm can reach 10ms per frame.
4. Conclusion
In this paper, for traffic target detection scenarios, considering the performance and speed of the algorithm, the YOLOV4 cropping model is used, and images of different weather and time periods at traffic intersections are used. 6 types of target detection models are trained, and the algorithm is deployed to edge computing. On the device, the model's mAP can reach 94.5%. Therefore, the YOLOV4 detection model is a detection algorithm that is very suitable for complex traffic scenes, which can meet the real-time performance of algorithm deployment and achieve better detection results.

Acknowledgments
This work was financially supported by fund project, that is, Young Talents in Higher Education of Guangdong, China, (No. 2019KQNCX231 and No. 2019KQNCX232).

References
[1] Cai Z, Fan Q, Fe Ris R S, et al. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection [C] European Conference on Computer Vision. Springer International Publishing, 2016.
[2] Fu C Y, Liu W, Ranga A, et al. DSSD: Deconvolutional Single Shot Detector [J]. 2017.
[3] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger [J]. IEEE, 2017:6517 - 6525.
[4] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. IEEE, 2016.
[5] Huang Z, Huang L, Y Gong, et al. Mask Scoring R-CNN[C] 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.
[6] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into High Quality Object Detection [J]. 2017.
[7] Bochkovskiy A, Wang C Y, Liao H. YOLOv4: Optimal Speed and Accuracy of Object Detection [J]. 2020.