Vehicle and pedestrian target detection in auto driving scene

Wanbo Yu 1, Pengjie Ren 1*

1 College of Information Engineering, Dalian University, Dalian 116622, China
*Corresponding author’s e-mail: yuwanbo@dlu.edu.cn

Abstract. To improve the target detection accuracy and speed of autonomous driving in various weather environments and small target traffic scenarios, an improved YOLOv4 target detection model based on CSPDarknet45_G backbone network is proposed in this paper. By adding a new DBG module which consists of DarknetConv2D + BN + GELU activation function, this model is enhanced in generalization ability and accuracy. We also improved Res unit residual module to enhance shallow features fusing with deep feathers and reduced the number of neurons in the CSP module to simplify the module structure. The K-Means++ clustering algorithm is introduced to obtain the size of the prior box used for target detection to satisfy the data set in this paper. In the captured target vehicle image data set, the model detection result shows that the improved YOLOv4 model achieves an average detection accuracy of 90.45%, a recall of 94.37%, and an FPS of 50 frames per second when the IOU is taken as 0.5, which meet the real-time and accuracy of the detection task in this paper.

1. Introduction

In the scenario of autonomous vehicle driving, the target detection of road obstacles during driving is a key task for autonomous driving. The current mainstream target detection frameworks can be divided into one-stage algorithms based on regression (SSD [1], YOLO [2]) and two-stage algorithms based on candidate region extraction (R-CNN, Faster R-CNN [3]). Automated vehicle target detection remains extremely challenging due to lighting, occlusion, complex road backgrounds, and unexpected conditions. The YOLOv4 model has been used for target detection in traffic scenes [4-6]. However, YOLOv4 model couldn’t reach a good balance between detection speed and detection accuracy, and it is easy to miss detection in small target detection. Besides, it cannot adapt to the detection tasks in different weather environments. In this paper, the CSPDarknet45_G network with shallower model depth is used as the backbone network of the improved YOLOv4 to improve the overall detection speed of the model. A new DBG module consisting of DarknetConv2D+BN+GELU activation function is added. We delete the last 1X1 DBM convolution layer of the CSP module resulting in a larger improvement in small target recognition accuracy. PCA colour enhancement is applied in the data set to improve the detection accuracy of the model under different weather conditions. Bilateral Filtering reduces the noise data and improves the training effect. The K-Means++ clustering algorithm is used to obtain the suitable size of Anchor in the detection target task, it effectively reduces the error of ground truth box and prior box and improves the detection accuracy of the model.

2. Improved YOLOv4 model

2.1. Add DBG module

In order to improve the nonlinearity and generalization of the model, it is necessary to add a stochastic regular (dropout), stochastic regular is quite different from nonlinear activity. The GELU [7] activation function is used to improve the nonlinearity of the model. By adding the DBG module, the model's ability to capture shallow features is enhanced, which is beneficial for detecting small targets. The DBG module consists of DarknetConv2D + BN + GELU activation function, which improves the model's generalization ability and accuracy.

2.2. Delete 1X1 DBM convolution layer

In this work, we delete the last 1X1 DBM convolution layer of the CSP module. This simplification leads to a large improvement in small target recognition accuracy. Bilateral Filtering reduces the noise data and improves the training effect. The K-Means++ clustering algorithm is used to obtain the suitable size of Anchor in the detection target task, it effectively reduces the error of ground truth box and prior box and improves the detection accuracy of the model.
function introduces the idea of stochastic regularity which is a probabilistic description of the neuronal input. GELU multiplies the input by an \( \Phi(x) \) and \( X \) obeys the standard normal distribution. When the input \( x \) decreases, the input will be dropout with a higher probability, the activation transformation will depend on the input randomly. The expression of GELU activation function is shown in formula (1).

\[
GELU(x) = x \Phi(X \leq x) = x \Phi(x)
\]  

Assuming it is a standard normal distribution, the mathematical formula for its approximate calculation is as equation (2).

\[
GELU(x) = 0.5x \left( 1 + \tanh \left( \sqrt{2 / \pi} \left( x + 0.044715x^3 \right) \right) \right)
\]

By adding the DBG module composed of DArknetConv2D + Bn + GELU activation function, the experimental effect is better than Leaky ReLU and Mish. Improved the ResNet and CSP structure on the basis of the DBG module. The improved ResNet\(^{[8]}\) performs a DBM convolution with a 1X1 convolution kernel on the input, then performs a DBG convolution operation with a 3X3 convolution kernel and a padding of 1, and finally superimposes the result on the previous input for addition. Improve the CSP structure, first perform a DBL convolution operation with a convolution kernel size of 1X1 and a step size of 2, and scale the input image size to 1/4 of the previous one to achieve the purpose of downsampling. Then a large stack of residuals is performed, where Res Unit *N indicates that N ResNet residual operations are performed in the network. The DBM module with a convolutional kernel size of 1X1 at the tail of the CSP structure is deleted to reduce the network depth of the CSP structure. The structure of the improved module is shown in Figure 1.

![Figure 1. Improved modules.](image)

2.2. **CSPDarknet45_G backbone network**
CSPDarknet45_G contains 5 CSP_ResBock_Body blocks, and each CSP block contains N Res unit residual modules. The number of Res units included in the initial CSP residual block to the last CSP residual block is 1, 4, 4, 6, and 4 respectively, and the number of convolutional layers included in each Res unit is 2. From the third CSP to the fifth CSP residual block, the corresponding prediction output feature map sizes are 76X76, 38X38, 19X19, corresponding to the detection of small, medium, and large targets respectively. Compared with the CSPDarknet53 backbone network model, the CSPDarknet45_G backbone network has a shallower depth and a faster detection speed, which can better retain local and detailed features and reduce the loss of image semantics. The backbone network structure is shown in Figure 2.

2.3. **K-Means++ clustering algorithm**
Aiming at the instability of the Anchor Box selected by the K-Means algorithm, this paper uses the K-Means++ clustering algorithm to select the size of the Anchor Box suitable for the model. The improved part is that the initial 9 a prior box cluster centers are selected by calculation. The initial Anchor Box clustering center selected by the K-Means++ clustering algorithm is more stable and more reasonable without too much fluctuation. The K-Means++ clustering algorithm is as follows:
Step1: Randomly select one from the labeled data set as the initial Anchor Box cluster center.

Step2: Calculate the distance D(x) between each real box in the data set and the nearest Anchor Box cluster center, and classify it into the corresponding Anchor Box category. Then calculate the probability that each ground truth box is selected as the next Anchor Box cluster center, and select the ground truth box with high probability as the next cluster center. The probability expression is as in formula (3), where D(x) is the distance from the ground truth box to the current closest Anchor Box cluster center.

$$p = \frac{D(x)^2}{\sum_{i=1}^{n} D(x_i)^2}$$

(3)

Step3: Repeat Step2 until 9 initial Anchor Box clustering centers are selected.

The feature map sizes obtained on the training set through the K-Means++ clustering algorithm are: ((8x17), (14x30), (20x46); (8x17), (14x30), (20x46); (85x92), (162x147), (296x331);

3. Experimental results and analysis

The data used in this paper is a data set of pictures taken by real traffic roads. The total number of images in the training set is 4,531, the total number of images in the verification data set is 1,365, and the total number of images in the test set is 2,357.

3.1. Model improvement ablation experiment comparison

The original YOLOv4 model is represented by A, the models B, C, D, and E are successively improved on the basis of the previous model, and the model F is to independently verify the influence of the CSPDarknet45_G backbone network on the performance of the model. Models G and H
perform PCA and Gaussian bilateral fuzzy filter data enhancement processing in turn. For example, as shown in Table 1.

Table 1. Ablation comparison experiment

| Models       | YOLOV4 improvement mAP% | Recall% | FPS |
|--------------|-------------------------|---------|-----|
| A            | YOLOv4                  | 87.73   | 91.52 | 48  |
| B            | A+GELU                  | 88.75   | 92.38 | 43  |
| C            | B+CSP                   | 89.03   | 92.60 | 45  |
| D            | C+CSPDarknet45_G        | 88.79   | 92.41 | 51  |
| E            | D+ K-Means++            | 90.04   | 93.72 | 51  |
| F            | A+CSPDarknet45_G        | 85.14   | 89.15 | 54  |
| G            | E+PCA                   | 90.31   | 94.06 | 51  |
| H            | G+ Bilateral Filtering  | 90.45   | 94.37 | 50  |

The improved YOLOv4 network compared to the YOLOv4, YOLOv3 and Faster RCNN models in each category AP comparison is shown in Table 2.

Table 2. Detection accuracy of different categories (unit:%)

| Category        | Improved YOLOv4 mAP% | YOLOv4 mAP% | YOLOv3 mAP% | Faster RCNN mAP% |
|-----------------|----------------------|-------------|-------------|------------------|
| Pedestrian      | 94.5                 | 91.4        | 85.3        | 92.5             |
| Bicycles        | 87.9                 | 85.6        | 80.2        | 86.2             |
| Motorcycles     | 92.3                 | 90.9        | 83.1        | 91.1             |
| Cars            | 96.9                 | 95.7        | 93.6        | 95.3             |
| Red Light       | 89.9                 | 85.3        | 78.5        | 82.5             |
| Green Light     | 87.7                 | 83.5        | 75.0        | 81.7             |
| Yellow light    | 84.1                 | 81.7        | 74.3        | 79.6             |

The comprehensive performance comparison of the four models is shown in Table 3.

Table 3. Comparison of comprehensive model performance

| Models          | mAP%  | Recall% | FPS |
|-----------------|-------|---------|-----|
| Improved YOLOv4 | 90.45 | 94.37   | 50  |
| YOLOv4          | 87.73 | 91.52   | 48  |
| YOLOv3          | 81.42 | 87.69   | 53  |
| Faster RCNN     | 86.98 | 93.15   | 19  |

The actual road detection effect of the improved YOLOv4 model is shown in Figure 3.
4. Conclusions
In this paper, a CSPDarknet45-G backbone network vehicle and pedestrian target detection model is given based on YOLOV4. In the proposed model, we added a new DBG module composed of GELU activation function, improved Res Unit residual network, improved CSP network structure, and reduced CSP structure depth and parameter amount. The K-Means++ clustering algorithm is introduced to obtain a suitable prior frame, which improves the accuracy of model detection. Experimental results show that the improved YOLOv4 model can significantly improve the detection effect of small targets such as traffic lights and pedestrians in actual traffic scenes. The improved model mAP reaches 90.45%, Recall reaches 94.37%, and FPS reaches 50 frames per second. The overall performance is higher than other network models, and it can well take into account the detection speed and detection accuracy.

Acknowledgments
First and foremost, I would like to show my deepest gratitude to my supervisor, Professor. Yu Wanbo, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance in every stage of the writing of this thesis. Without his enlightening instruction, impressive kindness and patience, I could not have completed my thesis. Last but not least, I'd like to thank all my friends, especially my three lovely roommates, for their encouragement and support.

References
[1] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.
[3] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.
[4] XN X k, Ma Y, QIAN X, ZHANG H. et al. Scale-aware real-time pedestrian detection for autonomous driving scenarios[J]. Chinese Journal of Graphics, 2021, 26(01):93-100.
[5] GU L P, SUN S Y, LI X, LIU X H, SONG Q Q, Multi-objective vehicle and pedestrian tracking algorithms for unmanned vehicle driving scenarios[J]. Small Microcomputer Systems, 2021, 42(03):542-549.

[6] LI Y P, HOU LING Y, WANG C. Motion target detection in autonomous driving based on YOLOv3[J]. Computer Engineering and Design, 2019, 40(04):1139-1144.

[7] Hendrycks D, Gimpel K. Gaussian error linear units (gelus)[J]. arXiv preprint arXiv:1606.08415, 2016

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.