International Journal of Robotics and Automation, Vol. 36, 2021

Special Issue: Artificial Intelligence in Environmental Automation Systems

VEHICLE TYPE DETECTION BASED ON RETINANET WITH ADAPTIVE LEARNING RATE ATTENUATION

Yiliu Xu,* Peng He,* Hui Wang,* Ting Dong,* and Pan Shao*

Abstract

Aiming at the problem that the accuracy and speed of the current target detection algorithm cannot be balanced, this paper uses RetinaNet as the basic framework for vehicle type detection and proposes an adaptive learning rate attenuation (ALRA) on the basis of the least squares, which can promote model convergence effectively. The experimental results show that training time cost is reduced and accuracy is improved by using ALRA to train the model.

Key Words

Learning rate, RetinaNet, focus loss, vehicle type detection, pyramid feature network

1. Introduction

With the rapid growth of current computer hardware performance, deep convolutional neural networks have shown superior advantages over traditional algorithms in areas such as image classification and image object detection. Deep neural network is a supervised learning method. On the basis of a large number of labelled samples as training sets, the weight parameters are updated and the objective function is fitted. Since a large number of parameters are introduced, the biological neural network is simulated on the structure, so the depth model phase compared with the traditional algorithm, it has great advantages. In the field of image recognition, the convolutional neural network model constantly refreshes the accuracy rate in the ImageNet contest, and it rides in the field [1].

The training of convolutional neural networks is actually a process of minimizing loss. To make the model fit the objective function better, in the process of neural network training, the model calculates the cross-entropy loss between the predicted distribution and the real distribution (ground truth), and finds the partial derivative of the activation function to carry out the back propagation for updating weight parameter. The magnitude of the weight update is expressed by the learning rate [2]. In the initial training period, because the model can hardly fit any sample, the learning rate should be set larger so that the model converges as soon as possible [3], but when the loss is reduced to a certain extent, if the learning rate stops changing, it will cause that the loss shocks repeatedly and cannot converge [4]. This situation is caused by the fact that the learning rate is relatively high and cannot converge to the local optimal solution. Therefore, it is important to formulate the corresponding learning rate attenuation strategy [5].

In the field of deep learning image processing, target detection is a hot research direction. Target detection needs to locate and classify objects in the image at the same time. The corresponding mainstream algorithms can be divided into two categories, one is the one-stage detection algorithm represented by YOLO [6] and SSD [7], and the other is the two-stage detection algorithm represented by the RCNN [8] series of algorithms, in terms of speed, because of directly regression and classified by the one-step detection method such as YOLO and SSD, and the RCNN series algorithm first uses the candidate network to extract the candidate frame, then use a relatively simple sub-classification to get results, it costs more time, so one-stage detection algorithm has a clear advantage in speed. In terms of accuracy, the RCNN series two-step detection method fixes the proportion and number of positive and negative samples when generating positive and negative samples, and uses the RPN network to provide candidate region frames for the detection network, while the YOLO and SSD algorithms have no candidate networks. The use of point-by-point sampling produces positive and negative samples, resulting in an extremely imbalanced number of positive and negative samples, so the accuracy is lower relative to the RCNN series algorithm. To meet the requirements of accuracy and speed at the same time, He *et al.* proposed a target detection network, named RetinaNet [9] on the basis of focal loss. The RetinaNet

* College of Computer and Information, China Three Gorges University, Yichang 443002, China; e-mail: {549342071, 2570076609}@qq.com, {hpeng, dongt}@ctgu.edu.cn, panshao@whu.edu.cn

Corresponding author: Peng He

Recommended by Dr. Dong Ren

(DOI: 10.2316/J.2021.206-0625)

network consists of ResNet, feature pyramid networks (FPN) and full convolution neural network (FCN), and aims at the problem of too large total loss proportion of easy-to-sort samples, which causes difficult samples cannot to be effectively trained. On the basis of the traditional cross-entropy loss function [10], a focal loss function is proposed, the characteristics of this loss function are very clear, for easy-to-sort samples, the loss is greatly attenuated by adding weights; for the difficult-to-separate samples, the loss is slightly attenuated. Because the loss ratio of the easily-divided samples is reduced, the learning ability of the model for the difficult-to-separate samples is greatly improved. With using a one-step detection architecture, RetinaNet is comparable in speed with YOLO and SSD, and even exceeds the two-step detection method with the highest accuracy in the same period, which combines the advantages of high precision and high speed.

In the field of spectral analysis, each material has its own unique characteristic line. Therefore, support vector machine (SVM) and random forest are commonly used in the field of spectral analysis [11]. According to this feature, the recognition model constructed by deep learning also achieves good results. In the field of multi-spectral remote sensing image processing and geological hazard warning [12], the use of CNN model to classify multi-spectral images has also achieved the effect of surpassing traditional algorithms such as SVM, and it has broad application prospects in the field of remote sensing.

Traffic vehicle type detection is a hot direction for deep learning target detection applications [13]. In forestry, to solve the illegal logging, it is necessary to accurately identify some wood-loaded vehicles in real time. This article takes this as the background, on the basis of RetinaNet. The framework proposes an adaptive learning rate attenuation (ALRA) on the basis of least square method for the problem of model loss cannot decline with improper learning rate setting in model training. On the basis of 5,634 traffic cameras, the sample set for photo production is trained and tested. The vehicle type is divided into canvas, mask canvas, others buses, others empty cars, others minibuses and wood.

2. Detection Method

2.1 Relative Work

Takase *et al.* [14] proposed a neural network training method on the basis of training loss, and creatively proposed a loss-based learning rate update strategy. Experimental results show that this ALRA algorithm can effectively improve the test accuracy when training neural networks. Xu *et al.* [15] proposed a structure with multi-branch residual network and solved the problem of neural network gradient disappearance. The accuracy of the test improves by using the adaptive cosine learning rate algorithm. Yamada *et al.* given a method of adaptively selecting the learning rate (lr) using short-term pre-training, which reduces the computation time and improves the efficiency of finding the appropriate rate (lr) [16]. Blier *et al.* [17] proposed an ALRA learning rate update algorithm, in which each neuron obtains its own learning rate from a cross-order feature distribution because all neurons are looking for their own learning rate and thus compared with stochastic gradient descent (SGD), it has the best learning rate.

2.2 Structure of RetinaNet

He *et al.* found that the reason why the one-step detection method is lower in accuracy than the two-step detection method is mainly caused by the imbalance quantity between positive and negative samples of the one-step detection method. This directly leads to the loss of the model being dominated by the absolute number of easily separable samples. To solve this problem, a new loss function Focal Loss is proposed, and RetinaNet is designed on the basis of this loss function. Focal loss is constructed by adding a weight attenuation term $\alpha(1-p_t)\gamma$ on the cross-entropy loss function, which causes the loss of the easily-divided sample to be greatly attenuated and the loss of the hard-divided sample is weakly attenuated, it greatly enhances the learning ability of the model for difficult samples. The experimental results of He *et al.*, while $\alpha=0.25$, $\gamma=2$, RetinaNet on the basis of the focal loss compared with the conventional one detection method, the precision is greatly improved, which proves the validity of focal loss. The traditional cross-entropy (CE) formula is as follows, and represents the probability that the sample is correctly classified.

$$CE = -\lg(p_t) \tag{1}$$

Focal loss function formula:

$$FL = -\alpha(1-p_t)^{\gamma} log(p_t) \tag{2}$$

On the basis of this new loss function, the author designs a one-step detection framework named RetinaNet. As Fig. 1 shows, RetinaNet uses ResNet and FPN as backbone networks, draws lessons from anchors of the two detection methods *faster RCNN* [18], determines positive and negative samples with IoU by setting rectangular frames of different scales, and finally uses two FCN subnetworks to solve classify and regression tasks. Because of the direct classification and regression of one-step detection framework, there is no candidate frame mechanism, and

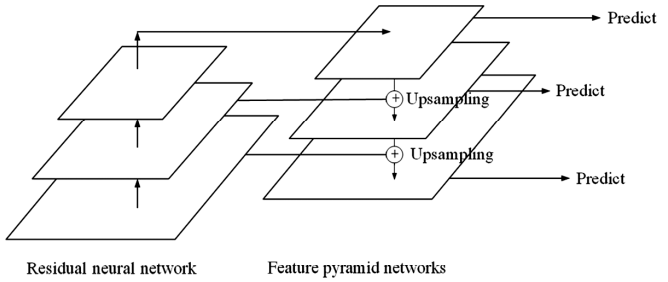

Figure 1. Schematic diagram of RetinaNet structure.

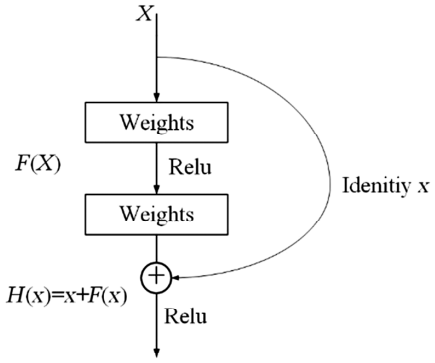

Figure 2. Schematic diagram of the residual module.

focal loss is used to balance the loss ratio of positive and negative samples. RetinaNet has obvious advantages over the conventional one-step and two-step detection methods in speed and accuracy.

2.3 Residual Neural Network

For the traditional AlexNet [19], VGG series [20], GoogleNet [21] and other networks which just expand their depth and width of layer, due to the gradual increase of depth and the complexity of the network, a large number of partial derivatives accumulate in the back propagation of the network, resulting in gradient explosion and gradient disappearance, so the increase of depth will directly lead to the model cannot converge, and the accuracy will even decrease. To solve this problem, He *et al.* proposed a ResNet network model on the basis of residual module. The structure of the residual module is shown in Fig. 2. For each remaining module, the input x is added directly to the output $F(x)$, with $H(x) = x + F(x)$ as the total output, the training process of the model can be regarded as a process of learning identity mapping $x = F(x)$. Because the residual module is added to the ResNet, the identity mapping learned by the model becomes $F(x) = H(x) - x$, from direct training $F(x) = x$ to training $F(x) = 0$, because the input is added directly to the output. Therefore, the gradient disappearance problem is effectively solved in reverse transmission, so that deeper network training can be carried out.

2.4 Feature Pyramid Networks

FPN is a network with multi-scale features. In traditional convolution neural network, with the increase of depth, the scale of feature map is smaller and more shallow semantic information is lost. Especially, the small object on high-layer map, whose feature has been very weak. For this problem, the FPN network is constructed with the addition of up-sampling and the same order forward characteristic map to construct the pyramid characteristic network. Due to the output of semantic information of different scales, FPN can be greatly enhanced the detection effect of the algorithm on different scale targets [22].

After the input box is generated by ResNet [23] and FPN, two FCN sub-networks are used for regression and classifying the input box. The two FCN have the same structure, the parameters are not shared and the feature scale is not changed.

2.5 Adaptive Learning Rate Attenuation Strategy

The learning rate represents the amplitude of the model training, too high learning rate will lead to the model cannot converge to the optimal solution and concussing around optimum solution iteratively, and too small learning rate will lead to the model convergence speed is too slow, the training time will be greatly increased [13]. Therefore, it is of great significance to set the appropriate learning rate in different training periods of the model to promote the model to converge to the optimal solution as soon as possible. The traditional learning rate attenuation strategy adopts exponential attenuation strategy. When a certain number of training iterations is reached, the learning rate attenuation rate decays to 10% of the previous iterations. This learning rate attenuation strategy can promote the model to converge to the global optimal solution to a certain extent [14], but it does not combine the trend of loss concussion. Loss concussion of the model objectively shows that with the increase of the number of training wheels, the loss does not decrease and even rises [15], so the attenuation learning rate combined with the trend of loss concussion is considered [16]. Assuming that the learning rate decreases when training iterations pass n iterations, the least square method is used to fit the discrete n loss points, and the first-order linear function $y = \beta x + \delta$ is used as the fitting target to determine the loss shock coefficient β and δ. The attenuation amplitude of the learning rate is determined according to β. When $\beta > -0.1$, the learning rate is 10% of itself, and no change in other cases. The specific learning rate (lr) update formula is as follows, lr_{ori} representing the original learning rate attenuation strategy and lr_{ours} representing the strategy proposed in this paper.

$$\begin{aligned} lr_{ori} &= 0.1lr, iters \\ &= n \times k(k = 1, 2....), \quad lr_{ours} = \begin{cases} 0.1 \times lr, \ \beta > -0.1 \\ lr, others \end{cases} \end{aligned} \tag{3}$$

The formula shows that when the loss decreases by <0.1 or the loss shows an upward trend, the learning rate decays to 10% of itself. In other cases, the learning rate remains unchanged. Due to the guiding role of the trend of loss changes, the model can be closer to the optimal solution under the fixed training iterations.

As Fig. 3 shows, in the process of model training iteration, this paper dynamically adjusts the learning rate according to the ALRA strategy. In the training process of RetinaNet, the network parameters are updated by the loss function back-derivation. Each updating is recorded as an iteration. The training process updates itself by add a fixed weight on the learning rate after n iterations. After adding the ALRA strategy in this paper, the fixed weight becomes a variable whose value is calculated by the strategy proposed in this paper.

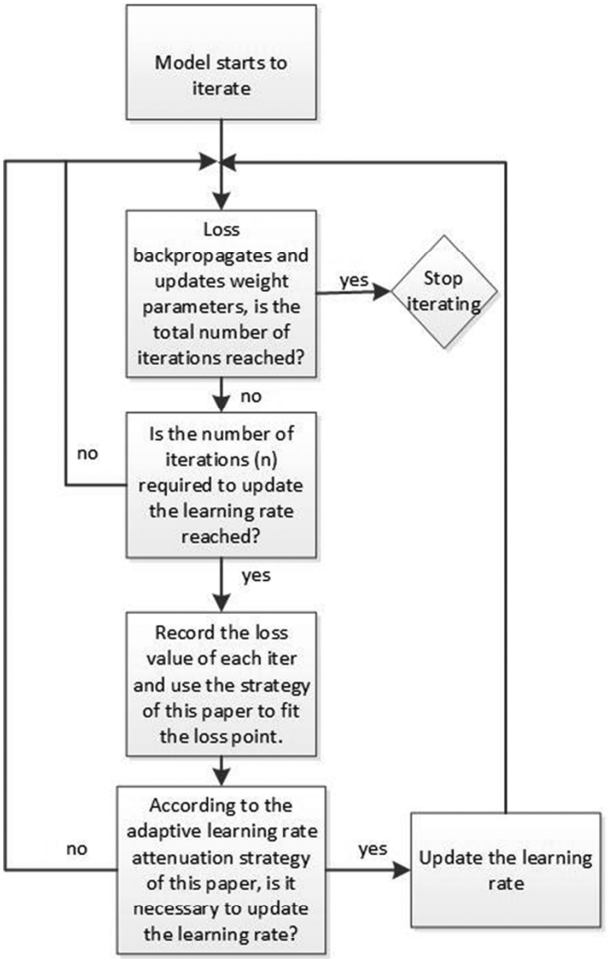

Figure 3. Updating process of adaptive learning rate.

3. Experiment

3.1 Dataset

The dataset in this paper is made from 5,634 photos taken by traffic cameras. First of all, six types of targets are identified, including canvas, mask canvas, others buses, others empty cars, others minibuses and wood. For the original jpg format picture, the corresponding vehicle type in the picture is framed by label and the xml file is generated, and then it is divided into training set and test by using python script, written to the txt document used for training, and finally made as a standard VOC format for training and testing. The Table 1 is the distribution of vehicles.

3.2 Experimental Platform

The hardware platform of this experiment is CPU of core i7-3330, GPU of GTX1060, programming language is python3.5, using Google's deep learning framework tensorflow1.12.0 and keras2.2.4 and using CUDA9.0 and CUDNN7.3.1 for GPU model training acceleration. VC build tools are used to compile the python part of the project, and pycharm is used as python interactive interface.

Table 1
Number of Vehicles of Various Types

	Canvas	Mask Canvas	Others Buses	Others Empty Cars	Others Minibuses	Wood
Count	926	1550	1268	1624	4056	422

Table 2
Training Efficiency Comparison

Experimental Algorithm	The Number of Iteration Required to Reduce the Loss to Below h ($n = 2,000$, $h = 10$)	Training Time (Hour)
faster RCNN	$31 \times 2{,}000$	29.5
RetinaNet	$30 \times 2{,}000$	28.3
RetinaNet + ALRA	$\mathbf{27 \times 2{,}000}$	**26.4**

3.3 Model Training and Evaluation Index

To minimize the training time costs of the model as much as possible, this paper uses transfer learning to train the model. Migration learning is a process of initializing new task network parameters directly by using model parameters trained on large datasets. By fixing the network parameters of bottleneck layer, a new network layer is fine-tuned at the end of the network for new tasks. Migration learning greatly reduces the parameters that need to be updated in the model, so it is widely used in model training. The feature extraction network used in this paper is resnet101, the deep residual network with depth of 101 layers. Therefore, the resnet101 pretrained on imageNet is used to be foundation of training, using the model file to initialize some parameters and fix them to bottleneck layer in the initial stage of model training, and then sets up a new network layer and trains to complete the new classification and regression tasks.

The specific training steps can be summarized as follows:

1. The sample is input into the ResNet feature extraction network to extract the deep features.
2. By sampling the deep features and adding the forward propagation features of the same order, a pyramid feature network is constructed in turn.
3. The region of interest (ROI) positive and negative samples are generated by setting the IoU threshold at each scale of the pyramid network. IoU > 0.7 is a positive sample and IoU < 0.3 is a negative sample.
4. Put the ROI into the sub-network for the training of classifiers and regression devices, using focal loss to significantly reduce the proportion of loss of easily divided samples.
5. Save the model and test it.

In the field of image recognition, recall and precision are usually used as evaluation indexes, but because these two indexes are affected by confidence threshold, the recall and precision of the same model will change under different confidence threshold. Therefore, the average precision (AP) is used to evaluate the detection effect of a single category, while the average precision mean (mAP) is used to evaluate the detection effect of the whole model. TP, the number of positive samples which is predicted as a positive sample, FN, the number of positive samples which is predicted as negative samples and FP, the number of negative samples which is predicted as positive samples. Assuming that there are k classes, using dense sampling draws the P–R curve, takes m confidence threshold evenly between [0, 1] and obtains the m pair P–R value, then the calculation formula of P, R, AP and mAP is as follows:

$$\begin{aligned} P &= \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, \\ AP &= \frac{p_1 + p_2 + \ldots + p_n}{n}, \\ \mathrm{m}AP &= \frac{AP_1 + AP_2 + \ldots + AP_k}{k} \end{aligned} \tag{4}$$

3.4 Results of Experiment

In this paper, the traffic vehicle dataset is used to train the model. The number of training iteration in all experiments is fixed at 70,000 iterations, n is 2,000, the initial learning rate is 0.001, the batch size parameter is set to 32 and the ratio of training samples to test samples is 1:1. To compare, three kinds of detection models are trained, including *faster RCNN*, RetinaNet and RetinaNet+ALRA, using 101-layer deep residual network as feature extraction network.

3.4.1 Training Efficiency Comparison

The main function of the ALRA strategy proposed in this paper is to shorten the training rounds required for the model training process. In the loss function, it represents that ALRA will spend less iteration for training. On the basis of experimental experience, in this experiment, when the average value of the n loss values calculated by the k time is less than a certain value h, the current iteration is recorded as $k \times n$. In our paper, n is 2,000 and h is 10. The total training iteration is 70,000.

As shown in Table 2, the bold value represents the best in the peer or column, the proposed method significantly reduces the training rounds and time required to reduce the model loss to <10, compared with the *faster RCNN* and the original RetinaNet, the average number of training rounds is reduced by 12.9% and 10.0%, the training time decreased by 4% and 6.7%, respectively, indicating that the method of this paper can effectively promote the convergence of the model and save time cost.

3.4.2 Comparison of Model Accuracy and Speed

In the test stage, setting different confidence threshold will lead to different results. 53,000 groups of recall-precision key pairs were obtained by dense sampling of confidence threshold, and the P–R curve was drawn with these key value pairs as horizontal and longitudinal coordinates, as shown in Fig. 4. The average precision mean (AP) of different types of vehicles and the mAP, experimental results of the model are compared with different models as shown in Table 3.

As Fig. 4 and Table 3 shows, the bold value represents the best in the peer or column, the abscissa of the P–R curve represents the recall rate, and the ordinate represents the accuracy rate. As the confidence threshold decreases, the recall rate increases and the accuracy decreases. For the overall evaluation model to identify the specific category, the average of all the accuracy is used. The value AP is used as an evaluation index to evaluate the test effect of the model on the category. In Fig. 4, compared with the *faster RCNN* model and the original RetinaNet model, the model trained by RetinaNet using the ALRA strategy has an upward trend in the four types of *canvas, mask canvas*, *others empty cars* and *others minibuses*. Compared with the model trained by *faster RCNN*, the P–R curve of this model is significantly improved on the *wood* model, but compared with the original RetinaNet, the P–R curve of the *wood* model has not changed significantly. The analysis may be due to the strong learning ability of the model on the *wood* model. It can reach the vicinity of the optimal solution at an early stage under the

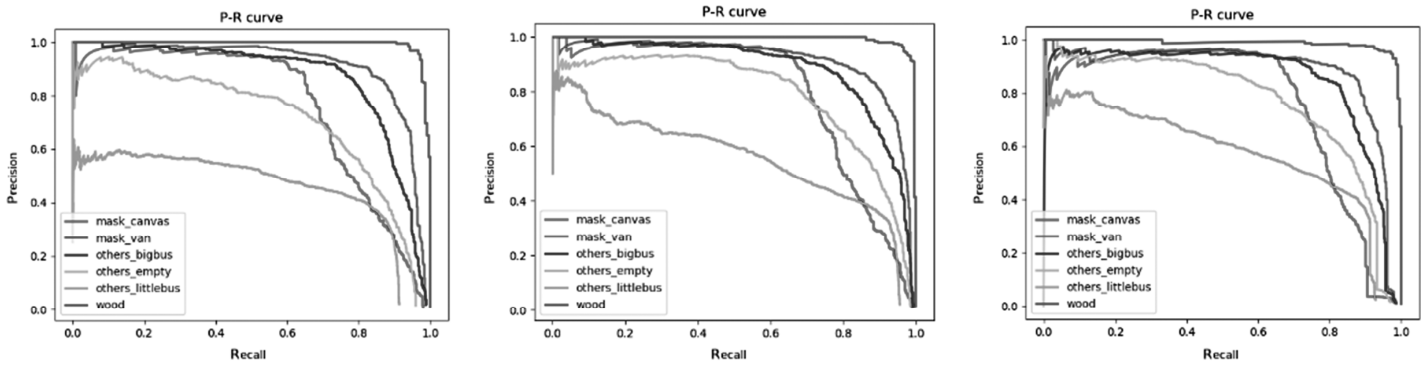

Figure 4. P–R curve of each model. (a) P–R curve of faster RCNN. (b) P–R curve of RetinaNet. (c) P–R curve of RetinaNet + ALRA.

Table 3
Accuracy Comparison of Each Model After the Same Number of Training Iteration

Algorithm	AP						mAP (%)	Speed (ms)
	Canvas	Mask Canvas	Others Buses	Others Empty Cars	Others Minibuses	Wood		
faster RCNN	0.719	0.851	0.820	0.748	0.585	0.934	77.6	196
RetinaNet	0.671	0.860	**0.870**	0.750	0.588	**0.978**	78.6	75
Ours RetinaNet	**0.735**	**0.871**	0.864	**0.758**	**0.616**	0.977	**80.4**	75

Figure 5. Vehicle detection effect in multiple environments during the day.

same number of iterations. On *others bus*, the RetinaNet model of this paper has a significant improvement on the P–R curve compared with the *faster RCNN* model, but it has a slight downward trend compared with the original RetinaNet model. The reason may be that the total loss of the model is dominated by other vehicle classes, which causes the fitting speed of this model is slower.

The experimental results show that in the model trained in this paper, due to the use of one-stage framework and focal loss as the loss function, the RetinaNet network outperforms the two detection methods in speed and accuracy. In terms of speed, the detection time of RetinaNet is reduced by 61.7% compared with *faster RCNN* on a single picture, and the average accuracy of RetinaNet is also slightly improved compared with *faster RCNN*. At the same time, the ALRA strategy proposed in this paper combined with RetinaNet is further used to train in the same way. In the case of the same number of wheels, compared with the original RetinaNet model, our model with ALRA is 1.8% higher on mAP, which shows that the ALRA strategy proposed in this paper can effectively promote the convergence of the model and improve the detection effect of the model when the number of training iteration is fixed.

3.5 Comparison of Detection Effects under Different Traffic Environment

To verify the robustness and generalization of the improved RetinaNet model to vehicle detection in different traffic environments, different road conditions, different lighting, different shooting distances and different weather conditions are extracted from the test set pictures as model inputs to detect the vehicles.

As shown in Figs. 5 and 6, the multi-environment pictures taken during the day, including a variety of traffic camera photos taken under different conditions. From the detection effect, most of the vehicles in the photos are correctly classified and have a high confidence level, all of which are above 0.9, indicating that the model has a good detection ability. According to the photos taken under different conditions, this model can detect the vehicle types correctly. It is shown that the model has strong robustness and generalization. Under the multi-environment detection of night photos, the model also shows excellent detection effect, which can accurately identify the vehicle types and has a high confidence level. Detection effect of the day and nights shows that the model in this paper can adapt to most of the vehicle target detection

Figure 6. Vehicle detection effect under nighttime multi-environment.

Table 4
Algorithm Comparison

Algorithm	mAP50 (%)	Speed (ms)
Yolo9000	66.5	43
faster RCNN	72.3	193
SSD	69.4	68
Ours RetinaNet	75.3	75

tasks under different shooting conditions and has strong portability.

3.6 Comparison of Effects on PASCAL VOC2007 Dataset

To further verify the effectiveness of this model, yolo9000, *faster RCNN*, SSD and RetinaNet models are trained on PASCAL VOC2007 dataset and tested with mAP50 as evaluation index, trained a total of 70,000 iterations, the experimental results are shown in Table 4.

The data in Table 4 shows that the RetinaNet model in this paper has obvious advantages over the other three algorithms, which exceeds the two-step detection method of *faster* RCNN on mAP and speed, compared with SSD and yolo9000, the mAP is significantly improved while the speed is similar, which shows that the RetinaNet model in this paper can take care speed and accuracy. Simultaneously, while it is compared with other algorithm of object detection.

4. Conclusion

In this paper, RetinaNet is used as the basic framework of vehicle detection model. To promote the model to converge to the optimal solution as much as possible under the same number of training iteration, an ALRA strategy on the basis of least square method is proposed. Compared with the original RetinaNet and *faster* RCNN algorithm training models, the mAP is significantly improved and cost of training time is reduced. Finally, traffic photos in a variety of environments are used as input of trained model, the vehicle recognition effect of the model in day and night environment is tested, and it is proved that the model has strong portability and robustness.

References

[1] C. Chen and F. Qi, Review on development of convolutional neural network and its application in computer vision, *Computer Science*, *46*(03), 2019, 63–73.

[2] Y. Ida, Y. Fujiwara, and S. Iwamura, Adaptive learning rate via covariance matrix based preconditioning for deep neural networks, *International Joint Conference on Artificial Intelligence*, Melbourne, VIC, Australia, 2017, 1923–1929.

[3] H. Zhao, F. Liu, H. Zhang, and Z. Liang, Research on a learning rate with energy index in deep learning, *Neural Networks*, *110*, 2019, 225–231.

[4] B. Chandra and R.K. Sharma, Deep learning with adaptive learning rate using Laplacian score, *Expert Systems with Applications*, *63*, 2016, 1–7.

[5] L. Fan, T. Zhang, X. Zhao, H. Wang, and M. Zheng, Deep topology network: A framework based on feedback adjustment learning rate for image classification, *Advanced Engineering Informatics*, *42*, 2019. https://doi.org/10.1016/j.aei.2019.100935.

[6] J. Redmon, S.K. Divvala, R.B. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, *International Conference on Computer Vision and Pattern Recogintion*, Las Vegas, NV, United states, 2016, 779–788.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, *et al.*, SSD: Single Shot MultiBox Detector, *European Conference on Computer Vision*, Amsterdam, Netherlands, 2016, 21–37.

[8] R.B. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *International Conference on Computer Vision and Pattern Recogintion*, Columbus, OH, United states, 2014, 580–587.

[9] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, and P. Dollár, Focal loss for dense object detection, *International Conference on Computer Vision*, Venice, Italy, 2017, 2999–3007.
[10] M. Altun and O. Pekcan, A modified approach to cross entropy method: Elitist stepped distribution algorithm, *Applied Soft Computing*, *58*, 2017, 756–769.
[11] D. Ren, C. Zhang, S. Ren, Z. Zhang, J.H. Wang, and A.X. Lu, An improved approach of cars for Longjing tea detection based on near infrared spectra, *International Journal of Robotics & Automation*, *33*(1), 2018, 97–103.
[12] P. Shao, W. Shi, and M. Hao, Indicator-Kriging-integrated evidence theory for unsupervised change detection in remotely sensed imagery, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(12), 2018, 4649–4663.
[13] A. Arinaldi, J.A. Pradana, and A.A. Gurusinga, Detection and classification of vehicles for traffic video analytics, *Procedia Computer Science*, *144*, 2018, 259–268.
[14] T. Takase, S. Oyama, and M. Kurihara, Effective neural network training with adaptive learning rate based on training loss, *Neural Networks*, *101*, 2018, 68–78.
[15] Y. Xu, H. Wang, X. Liu, and W. Sun's, An improved multi-branch residual network based on random multiplier and adaptive cosine learning rate method, *Journal of Visual Communication and Image Representation*, *59*, 2019, 363–370.
[16] K. Yamada, H. Mori, T. Youkawa, Y. Miyauchi, S. Izumi, M. Yoshimoto, *et al.*, Adaptive learning rate adjustment with short-term pre-training in data-parallel deep learning, *2018 IEEE International Workshop on Signal Processing Systems* (*SiPS*), Cape Town, South Africa, 2018, 100–105.
[17] L. Blier, P. Wolinski, and Y. Ollivier, Learning with random learning rates, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Wurzburg, Germany, 2019, 449–464.
[18] S. Ren, K. He, R.B. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine. Intelligence*, *39*(6), 2017, 1137–1149.
[19] A. Krizhevsky, I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Communications of the ACM*, *60*(6), 2017, 84–90.
[20] Y. Liu, B.C. Yin, J. Yu, and Z.F. Wang, Image classification based on convolutional neural networks with cross-level strategy, *Multimedia Tools and Applications*, 2017, 11065–11079.
[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, *et al.*, Going deeper with convolutions, *International Conference on Computer Vision and Pattern*, Boston, MA, 2015, 1–9.
[22] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, and S.J. Belongie, Feature pyramid networks for object detection, *International Conference on Computer Vision and Pattern*, Honolulu, HI, United states, 2017, 936–944.
[23] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *International Conference on Computer Vision and Pattern*, Las Vegas, NV, United states, 2016, 770–778.

Biographies

Yiliu Xu received his bachelor's degree from Hebei University of Science and Technology. He is now a graduate student of the Key Laboratory of Remote Sensing Monitoring and Analysis of Agricultural Environmental Safety in Hubei Province, China Three Gorges University. His main research interests are deep learning and object detection.

Peng He graduated from Xi'an Jiaotong University with a master's degree in computer software in 1989. He is an education information expert of Ministry of Education and a standing director of Education Information Technology, Hubei. His researches focus on network time synchronization and deep learning.

Hui Wang received her bachelor's degree in engineering from Shandong University. She is currently a graduate student at the School of Computer and Information, China Three Gorges University. Her main research interests are mobile traffic unloading and community testing.

Ting Dong received her Ph.D. from Wuhan University. She is currently a lecturer at the School of Computer and Information at the China Three Gorges University. Her main research direction is remote sensing image processing.

Pan Shao received his Ph.D. from Wuhan University. He is currently a lecturer at the School of Computer and Information at the China Three Gorges University. His main research directions are remote sensing image change detection and deep learning.