Multi-target detection method based on YOLOv4 convolutional neural network

Jinhui Liu1*

1 Department of Information, Beijing University of Technology, 100124, Beijing, China

*Corresponding author’s e-mail: tian1092852410@163.com

Abstract. In order to solve the problems of low detection accuracy, false detection and high miss detection rate of small targets in target detection tasks, this paper is a multi-target detection method based on YOLOv4 convolutional neural network. The proposed method is based on YOLOv4. The semantic information of high-level features is first propagated to the low-level network through FPN sampling, and then it is fused with the high-resolution information of the underlying features to improve the detection effect of small target detection objects. The information transmission path from the bottom to the top is enhanced by downsampling the feature pyramid, and finally the feature maps of different layers are fused to achieve relevant predictions. Experiments prove that the method proposed in this paper has good results.

1. Introduction

As an important research direction in the field of target detection, multi-target detection has always received widespread attention from researchers. At present, in-depth research has been produced in the fields of intelligent transportation, intelligent assisted driving and video surveillance [1]. Traditional pedestrian detection methods, such as HOG (histogram of oriented gradient) [2], DPM (deformable parts model) [3], and ACF (aggregate channel feature) [4], all use manual design or feature aggregation to obtain pedestrian features. With the major breakthrough of AlexNet [5] in image classification tasks in 2012, the use of convolutional neural networks (convolutional neural networks) to automatically learn the feature extraction process instead of traditional manual design is the current main research direction [6]. Target detection methods based on convolutional neural networks are mainly divided into two categories, one is a two-stage method, and the other is a single-stage method. The main idea of the first method is to use the cascade method [7] to further judge the category and position of the bounding box on the basis of generating the candidate target area. The other is a single-stage method. Take YOLO (youonly look once)[8] and SSD (single shot multibox detector)[9] as examples. The idea is to use a convolutional neural network to directly return the position and category. The introduction of convolutional neural networks as soon as possible has improved the performance of pedestrian detection algorithms, but the occlusion problem is still a major difficulty in pedestrian detection. Literature [10] uses a joint learning method to model different pedestrian occlusion patterns, but its detection framework is complex and cannot exhaust all situations. Literature [11] designed a new loss function to make the prediction frame keep close to the target real frame while keeping away from other real frames. This method is more flexible in processing occlusion and easier to implement. Literature [12] combines the aforementioned two ideas, and proposes a component occlusion sensing unit and an aggregation loss function to deal with the pedestrian
occlusion problem. Literature [13] deals with occlusion by introducing new supervision information (bounding box of pedestrian visible area). The idea is to use two branch networks to return to the pedestrian's whole body frame and the bounding box of the visible area respectively, and finally merge the results of the two branches. Improve detection performance. At present, target detection tasks mainly choose appropriate detection algorithms for different application scenarios: the single-stage algorithm has the fastest detection speed, but the accuracy is lower; the two-stage and multi-stage detection algorithms can obtain higher detection accuracy, but at the expense of detection speed.

This paper proposes a multi-target detection method based on YOLOv4 convolutional neural network. Based on the current best single-stage target detection algorithm YOLOv4, the semantic information of high-level features is first propagated to the low-level network through FPN sampling, and then it is combined with the bottom-level features. This method increases the information transmission path from the bottom to the top, and uses feature maps of different layers to fuse to achieve relevant predictions. This method is an end-to-end technology that effectively solves the problem of multi-target detection of static images, videos, and real-time videos.

2. Method

Our model architecture is shown in Figure 1. It consists of three parts: a front-end network for feature extraction, a feature fusion module, and a detection module for classification and regression operations. Adjust the size of the input image to 416×416 at the input, and input it to the network for training and detection. Our basic convolution block is a convolution layer that incorporates Batch Normalization (BN) and uses Mish and leakyRelu activation functions.

The front end of the model uses the backbone network composed of the CSPDarknet module stacked by the convolutional layer and the residual module, which effectively prevents the disappearance or explosion of the gradient on the basis of deepening the number of network layers to obtain a richer semantic information feature map, and In the backbone network, the dimensionality reduction of the feature map is achieved by 5 times of downsampling of the convolutional layer with a step size of 2 and a kernel size of 3; 2 times of upsampling are performed on the network neck and the shallow layer is realized with the PAN+SPP model structure The fusion of features and high-level semantic features and the fusion of multi-scale receptive fields make full use of the detailed features of the shallow network and improve the problem of feature loss of small targets; the detection head uses the idea of regression + classification to divide the input image separately It is 76×76, 38×38, and
19×19 grid images of three different sizes, which respectively realize the detection of small targets, medium targets and large targets.

The model adds the SPP module behind the backbone network, as shown in Figure 2. After the input feature map passes through a convolutional layer, it goes through three kernels of 5×5, 9×9, and 13×13 for maximum pooling, and then concat the obtained feature maps for channel splicing, and output. The number of channels becomes 4 times the original number of channels, and the feature map size remains unchanged. The output feature map size is:

$$output - size = \left( \frac{n + 2p - k}{s} \right) + 1$$

(1)

Where $n$ is the size of the input feature map, $p$ is the padding, and $s$ is the step size, which means rounding down. The SPP module obtains the receptive field information of the local area of the feature map and the receptive field close to the global by using the Maxpool layer with different sizes of kernels, and performs feature fusion. This operation of fusing different scales of receptive fields can effectively enrich the expression ability of feature maps, enhance the acceptance range of the output features of the backbone network, and separate important context information.

![Figure 2. SPP module.](image)

3. Experiment and result analysis

The experimental model training uses the hardware platform of Intel(R) Xeon(R) Gold 5218 CPU, GeForce RTX 2080 Ti 11G GPU. The software uses windows system, python 3.7, PyTorch1.5.0 deep learning framework. The experiment set the input image size to 416×416, the initial learning rate of the stage is 0.0001, the attenuation coefficient is 0.0005, the Adam gradient optimization algorithm is used, the batchsize is set to 4, and the training Epochs is set to 50 times. This experiment uses the data set COCO-2017. The data set contains 12 major categories and 80 minor categories. There are a total of 118287 images in the training set and 5000 images in the test set. Figure 3 is an example. This experiment uses mAP (Mean Average Precision) to evaluate network performance. The mAP index is used to evaluate multi-label image classification tasks and is an important index to measure the overall detection accuracy of the model in multi-category target detection.

$$mAP = \frac{\sum_{i=1}^{c} AP_i}{c}$$

(2)

It can be seen from Figure 4 that our model can well detect multiple targets in the picture, and at the same time gives the corresponding identification indicators. In addition, the position and size of the identification frame are determined from a visual angle, and the effect is just right. As shown in
Table 1, in terms of indicators, our method achieves a 70.2% mAP effect compared with the other two methods.

| Method            | mAP  |
|-------------------|------|
| scSE-YOLOv4-C     | 69%  |
| YoLov3            | 68%  |
| Our model         | 70.2%|

4. Conclusion
This paper proposes a multi-target detection method based on YOLOv4 convolutional neural network. We train convolutional network based on YOLOv4 on COCO2017 data set. Through experiments, we found that our proposed method has achieved good detection results on multi-target detection tasks, indicating that the method has certain practical value. The next step can be optimized to further improve the network and improve detection accuracy.

References
[1] Chen Y, Guo S, Zhang B, et al. A Pedestrian Detection and Tracking System Based on Video Processing Technology[C]// 2013 Fourth Global Congress on Intelligent Systems (GCIS). IEEE, 2014.
[2] Dalal N. Histograms of oriented gradients for human detection[J]. Proc of Cvpr, 2005.
[3] Felzenszwalb, Pedro F, Girshick, et al. Object Detection with Discriminatively Trained Part-Based Models.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010.
[4] Dollar P, Appel R, Belongie S, et al. Fast Feature Pyramids for Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8):1532-1545.
[5] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]// NIPS. Curran Associates Inc. 2012.

[6] Hashimoto J, Fukutomi M, Yamauchi H, et al. Pulse Radar Image Simulation with Ray-tracing Model for Underground Objection Detection[J]. IEICE technical report. Communication systems, 1998, 97.

[7] 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 & Machine Intelligence, 2017, 39(6):1137-1149.

[8] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[J]. 2015.

[9] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2016.

[10] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[J]. 2015:1904-1912.

[11] Wang X, Xiao T, Jiang Y, et al. Repulsion Loss: Detecting Pedestrians in a Crowd[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[12] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd[J]. 2018.

[13] Zhou C, Yuan J. Bi-box Regression for Pedestrian Detection and Occlusion Estimation[J]. 2018.