Study on Improvement of YOLOv3 Algorithm

Xinchao Liu1*, Haiyun Gan1 and Ying Yan1
1School of Automotive and Transportation, Tianjin University of Technology and Education, Tianjin 300222, China
*Corresponding author’s email: liuxinchao@tjzyjssfdh999.onexmail.com

Abstract. In order to optimize the problem of wrong detection and missed detection of small targets in complex environment, a target detection algorithm of YOLOv3-SPP5 was proposed. YOLOv3 in the deep learning algorithm has achieved excellent detection effect in target detection, but it is not perfect in the complex environment. In this paper, YOLO detection heads were added to the overall network of YOLOv3 to improve the extraction of multi-scale information features. In order to avoid the effect of feature extraction being reduced with the increase of network depth, Spatial Pyramid Pooling (SPP) were added to each YOLO detection head to optimize the extraction of deep network features. The optimized algorithm was named YOLOv3-SPP5. With the same setting of YOLOv3-SPP5 and YOLOv3, the experimental results on COCO data set show that the mAP of YOLOv3-SPP5 increases by 7.9, the Inference time and Volume do not increase significantly. The above experimental results show that the optimized YOLOv3-SPP5 algorithm is more suitable for target detection in complex scenes than YOLOv3 algorithm.

1. Introduction

With the development of computer vision technology, target detection algorithm has become an important part in the field of computer vision, and plays an extremely important role in unmanned driving, aerospace, road traffic and other fields. In recent years, target detection algorithms have become more and more mature, which makes it relatively easy to realize target detection based on intelligent system, and it has also been paid close attention in academia and industry. However, when detecting some small targets, multiple targets and targets in complex environment in the process of experiment, the frequency of missed detection and false detection will increase with the degree of complexity, which is an important difficulty and hot spot in the field of target detection. Because of the small target in the image or video pixels less differentiation is not obvious, but after the extraction of target detection algorithm and filtering can lead to the detection accuracy of sharp attenuation results thus resulting in fault inspection and leak detection, and in the complex environment of multi-target detection effect will be more obvious, so the small target of research is a very challenging job in target detection task.

Target detection is mainly to identify the position and category of obstacles from the image. As a great challenge in the field of object detection, the detection of small targets has quite high requirements for the mining of image resources and the detection of small-size features. Traditional target detection algorithm of small target detection has large limitation, in recent years, most of small target detection algorithm based on neural network optimization, mainly by changing the parameters of the network, image scaling, adding depth, reducing redundant network or in combination with some excellent design structure and modules to improve network performance in small target detection task. It contains a single-stage target detection algorithm based on SSD[1](DSSD, SSD, DSOD) series...
algorithm and YOLO\cite{2}(YOLO, YOLO9000\cite{3}, YOLOv3\cite{4}, YOLOv4) series algorithm, and a two-stage target detection algorithm based on Fast R-CNN\cite{5} and R-CNN\cite{6}. The single-stage target detection algorithm is faster than the two-stage target detection algorithm, but the average accuracy is not as high as the two-stage target detection algorithm. In order to maintain the real-time performance of target detection, the single-stage target detection algorithm based on YOLOv3 is preferred in this paper, and the COCO data set is selected as the learning object. Then, the detection accuracy of the network is enhanced by adding YOLO detection head and spatial pyramid.

2. YOLOv3

2.1. Algorithm introduction
YOLOv3 achieves the perfect balance of speed and accuracy by combining outstanding modules in the field of target detection. The YOLOv3 target detection algorithm mainly performs regression processing on obstacles. First, 416×416 color images are input into the network. After feature extraction, three feature images of different sizes are obtained: 13×13, 26×26, and 52×52. Then YOLOv3 divided the prediction graph into S×S cells. When the central coordinate of the target is located in a single cell, the cell is responsible for detecting the target. Each cell predicts the position information (x, y, w, h) and confidence of multiple bounding boxes. For COCO dataset, the dimension of the output prediction graph is 3×(80+5), and the final output feature graph sizes are 13×13×255, 26×26×255 and 52×52×255.

The backbone network of the model was obtained based on Darknet53. In the network, convolution kernels of 3×3 and 1×1 were mainly used. The 3×3 convolution kernel was used to extract the feature information on the image, while the 1×1 convolution kernel was mainly used to add nonlinear features and simplify the parameters of network training while the feature size remained unchanged. The CBL module represents a convolutional structure that is followed by BN normalization to reduce overfitting and selects the activation function Leaky RELU. Resn module uses the residual structure of deep residual network for reference, and Res block, a residual structure, can solve the problem of gradient disappearance in deep network training. In order to enhance the ability to extract network information, the input size of the network is adjusted to 608×608. The overall structure of the original YOLOv3 network model is as follows.

![YOLOv3 network model](image)

Figure 1. YOLOv3 network model.

2.2. Data sets
In order to verify the performance of the network, this experiment adopted COCO dataset. COCO data set with a large number of rich object types including people, vehicles, lights, parking piece, stop signs,
contains 200000 pieces of labeled images. the whole data set, the number of more than 1.5 million individuals, the inspection to optimize the performance of the network plays an important role. In this experiment, some pictures were selected for training and testing as shown in the following table.

| Table1. Data set information. |
|-------------------------------|
| Data set partition | The training set | The test set |
| graphics | 36000 | 9000 |

2.3. Network optimization

In the experiment, y4 and y5, two additional YOLO detection heads, were added to the YOLOv3 network in order to enhance the accuracy of YOLOv3 in detecting multi-scale target obstacles. In order to enhance the feature extraction of the deep network, SPP[7] modules were added in front of the YOLO detection head of YOLOv3 to improve the detection effect, which was combined into YOLOv3-SPP5. Since the small-scale feature map has a large receptive field, while the large-scale feature map has a small receptive field, the output dimensions of y1, y2, y3, y4 and y5 detection heads are 13×13×255, 26×26×255, 52×52×255, 104×104×255 and 208×208×255 respectively, which correspond to the detection of large target, medium target and small target respectively. As shown in the diagram below:

![Network structure diagram of YOLOv3-SPP5](image)

3. Experiment

In order to verify whether the detection effect of YOLOv3-SPP5 target detection algorithm has been significantly improved when detecting small targets, the following experiments were carried out.
3.1. experimental environment
The software environment of this experiment was built on the operating system Ubuntu16.04, and the deep learning framework was based on Darknet, which is dedicated to the YOLO series. The hardware configuration of the experimental environment is Inter Xeon E5-2603 CPU, with 16GB memory and NVIDIA P2000 GPU. The main software configuration is GPU acceleration libraries CUDA9.0, CUDNN6.0 and OpenCV3.0.

3.2 Network parameter configuration
SGD was used for network training, 100 epoch training was carried out on COCO data set, and 4 images were input into training each time. The initial learning rate was 0.001, multiplied by 0.1 at 70% and 90% of the total training process, respectively, and the momentum parameter was 0.9. Other parameters were carried out by default with the parameters of the original network.

3.3 Analysis of experimental results
Through experiments, the original YOLOv3 and the optimized YOLOv3 basic network were improved and the following experimental results were obtained. By comparing the following experimental results, the detection model with good detection performance was analyzed and selected.

Table 2. Experimental data.

| Model       | Input size | mAP  | Inference time(ms) | Volume (MB) |
|-------------|------------|------|--------------------|-------------|
| YOLOv3      | 416×416    | 55.30| 29.00              | 246.30      |
|             | 608×608    | 57.90| 51.00              |             |
| YOLOv3-SPP1 | 416×416    | 58.40| 29.70              | 250.70      |
|             | 608×608    | 60.50| 53.40              |             |
| YOLOv3-SPP3 | 416×416    | 59.40| 30.27              | 255.80      |
|             | 608×608    | 61.60| 54.52              |             |
| YOLOv3-SPP5 | 416×416    | 62.90| 32.96              | 260.10      |
|             | 608×608    | 65.80| 55.31              |             |

Figure 3. mAP comparison of YOLOv3, YOLOv3-SPP1, YOLOv3-SPP3 and YOLOv3-SPP5.

It can be seen from the experimental data in Table 1 that the MAP increases with the increase of the network input size, while the network inference time also increases greatly. Obviously, it is impossible to improve the detection accuracy by increasing the network input size. Therefore, a unique network structure, YOLOv3-SPP5, is designed to improve the performance of target detection. It can be seen from Figure 3 that whether the size of network input image is 416×416 or 608×608, the mAP shows an upward trend. There is no obvious growth trend in Inference time and Volume. Visualization effect analysis, from the detection effect figure 4 of YOLOv3 and the detection effect figure 5 of YOLOv3-SPP5, it can be clearly seen that there are missed traffic light and handbag in figure of the
detection result of YOLOv3.

Figure 4. Detection effect of YOLOv3.

Figure 5. Detection effect of YOLOv3-SPP5.

4. Conclusion
In this experiment, the YOLOv3 target detection algorithm was optimized and the existing YOLOv3 network with good detection performance was improved to meet the requirements of small target detection in complex environments. In this paper, COCO data set is adopted to enhance the diversity of samples, YOLO detection head is added to adapt to more small target detection scale, and SPP is added to the detection head of YOLOv3 to enhance the extraction of target pixels. The maximum value of MAP in the experiment is 65.80, which is 7.9 higher than that in the original YOLOv3 detection model under the same setting parameters. The optimization model still needs to be further optimized. Therefore, it is necessary to further optimize the model in the later work to further lighten the model on the premise of ensuring the detection accuracy.

Acknowledgments
This work was supported in part by the Tianjin University Discipline Leading Talent Training Program of China (No. SSW181030105) and the Research and Innovation Project for Postgraduates in Tianjin (Artificial Intelligence) (No. 2020YJSZXS32).

References
[1] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016) SSD: Single Shot MultiBox Detector. arXiv Preprint at https://arxiv.org/abs/1512.02325.
[2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015) You only look once: unified, real-time object detection. IEEE Computer Society.
[3] Redmon, J., Farhadi, A. (2017) Yolo9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6517-6525.
[4] Redmon, J., Farhadi, A. (2018) "YOLOv3: an incremental improvement." arXiv preprint at https://arxiv.org/abs/1804.02767(1804).
[5] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence. 39(6), 1137 – 1149.
[6] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Washington. pp. 580-587.
[7] He, K., Zhang, X., Ren, S., Sun, J. (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(9), 1904-1916.