A deep learning model S-Darknet suitable for small target detection

Miao Yang*, Xiangbin Shi

College of Computer Science, Shenyang Aerospace University, Shenyang, Liaoning Province, 110136, China

*Corresponding author’s e-mail: 182106074003@email.sau.edu.cn

Abstract. In order to solve the problem of low accuracy of small target detection in target detection, a small target detection model S-Darknet is proposed. The algorithm is designed based on the Darknet-53 network. First, a new backbone network is proposed, which fully extracts the feature of small objects and adapts to multi-scale detection. Then, in order to enhance the feature information of the target after the fusion, a feature enhancement module is added before each upsampling. Finally, the proposed algorithm was verified on the VOC2007, VOC2012 data sets and the actual data sets of railway freight locks. Experimental results show that this method has high detection accuracy under the premise of ensuring real-time performance.

1. Introduction

In recent years, deep learning has made significant progress in the field of computer vision, especially in the application of target detection. However, the accuracy of small target detection is often low. The proportion of small targets is small and the amount of information expressed by the pixels of the corresponding area is limited, which cause a lack of information. Therefore, how to improve the efficiency of small target detection is still a popular research content.

At present, small target detection algorithms are: Literature [1] uses an oversampling strategy. Pyramid Box [2] proposed a context-assisted method based on Anchor. SNIP [3] uses the scale normalization training scheme of the image pyramid. In addition, there are some improved algorithms for general algorithms. Common target detection algorithms can be divided into two categories. One is a two-stage target detection algorithm based on area recommendation. There are R-CNN [4], Fast R-CNN [5], Faster R-CNN [6], etc. This type of algorithm has higher accuracy but slower speed. The other is the One-Stage target detection algorithm based on regression. Representative algorithms include SSD [7], RFB[8], YOLO[9], YOLOv2[10], YOLOv3[11], YOLOv4[12].Their network layer propagation speed is fast and real-time.

In response to the above problems, this paper proposed the S-Darknet algorithm for small target detection. The first is to design a new backbone network based on the Darknet-53 network, which can fully extract the features of different scale targets and be better adapted to multi-scale detection. Then, in order to make the feature map more capable of expressing features during deep network processing and be more suitable for small target detection, a feature enhancement module (En Block) was proposed, the idea of which came from [13]. The method has achieved good results.

2. S-Darknet

The main reason for the difficulty of small target detection is that the relationship between the
semantic information of high-level features and the detailed information of low-level features is not fully utilized. In order to efficiently extract useful feature information and improve the accuracy of small target detection, this paper proposes the S-Darknet detection model.

2.1. Network framework
For small objects with fewer pixels, it is extremely necessary to extract more and effective information. However, DarkNet-53 feature extraction network only uses residual blocks to extract feature information. As the feature map is continuously down-sampled, the receptive field becomes larger, which is more suitable for the detection of large and medium-sized targets, and the loss of information leads to inadequate training for small objects. Moreover, the original Darknet-53 network adopts a structure similar to the top-down feature pyramid, which only introduces high semantic information from the deep into the shallow layer, without considering the auxiliary role of the shallow layer to the deep layer. And this structure makes the feature of small targets depend on the feature of large targets to a large extent. However, this dependence is not applicable to all scenarios. Therefore, this paper will improve the Darknet-53 network and propose a new network structure to be better suitable for feature extraction and detection. The overall structure of the S-Darknet network is shown in Figure 1.

![Figure 1. S-Darknet network structure.](image)

In Figure 1(b), the A network is an improved network similar to FPN in the original Darknet-53. It combines high-level semantics and low-level features to effectively transfer the features of small targets. The A structure first constructs a small-scale (high-level) feature map. Then it uses additional global average pooling (GAP) in the deepest layer of the backbone network to extract context Global information. GAP can learn richer semantic information, highlight the discriminative target area detected by the convolutional neural network, and spread more semantic information to a larger scale. The process of A structure is

\[ F_i^A = W_i^A \otimes (U(F_{i+1}^A) + C_i) \]

Where \( U \) is an upsampling operation with a step size of 2, \( W \) is a 3x3 convolution, and “+” is a normalization process of fusion. Because it iteratively propagates higher-level semantic information to more detailed lower-level feature maps, it will be better at detecting small objects.

The B structure is a fusion and split structure. Due to the sequential construction of the mapping relationship, the features constructed in the early stage always affect the subsequent features, and this...
dependency relationship will also cause the limitations of feature extraction. Therefore, the B structure first fuses the two highest layers into a feature map \( \alpha_s \) and the two lowest feature maps into \( \alpha_l \).

- \( \alpha_s = C_4 + U(C_5), \alpha_l = C_3 + D(C_2) \)

After the first round of combined features are obtained, the fusion of formula 3 and 4 is carried out.

- \( \beta_s = W_1^s \otimes \text{cat}(\alpha_s, D(\alpha_l)), \beta_l = W_2^s \otimes \text{cat}(\alpha_l, U(\alpha_s)) \)

Where \( W \) refers to a \( 3 \times 3 \) convolution operation, \( \text{cat} \) is a combination operation according to channel dimensions. \( \beta_s, \beta_l \) are feature maps that merge all layer information. Then adjust the size of them.

- \( F_2^B = U(\beta_l), F_3^B = \beta_l, F_4^B = \beta_s, F_5^B = D(\beta_s) \)

Through the above two rounds of fusion and splitting operations, the feature map obtained contains feature information from all levels. The B structure first fuses high-level features and low-level features, and decomposes the fused features into multiple scales, which can better detect small targets.

Finally, the structure A and B are integrated into one network to adapt to small targets of different scales. The internal structure of the backbone network is similar to a pyramid structure. The calculation cost is low, and no additional parameters are required to participate in the calculation.

- \( F_i = F_i^A + F_i^B, i = 1,2,3,4,5 \)

2.2. Feature enhancement

If the target is too small, information will be lost in both the up-sampling and down-sampling process. The small target may be ignored during training, resulting in insufficient training, thereby reducing the detection rate of the small target. Therefore, this paper proposes the feature enhancement module En Block to highlight the features of small targets and make small target training more adequate.

This module is used to enhance the target features during the training process. The idea is to keep the background part unchanged and highlight the marked part. The main process is shown in Figure 2.

![Figure 2. En Block.](image)

Its input is the activation tensor by the activation function of the previous layer. Each channel of the feature map is correspondingly added, and then averaged to obtain a single feature matrix \( b \). Then according to the information provided by groundtruth, the target position is set to 1, the background part is set to 0, and the matrix \( g \) is obtained. Finally, multiply the feature matrix \( b \) and \( g \) and add to each channel of the original feature map.

- \( b(i,j) = \frac{1}{d} \sum_{c=1}^{d} a_{(c,i,j)} \)
- \( g(i,j) = \begin{cases} 1, \text{bbox exists at cell}(i,j) \\ 0, \text{no bbox exists at cell}(i,j) \end{cases} \)
- \( a^{l+1}_{(c,i,j)} = a^l_{(c,i,j)} + b(i,j)g(i,j) \)

Where \( a^l \) and \( a^{l+1} \) refer to the activation tensor of \( l \) and \( l+1 \) layers. \( d \) refers to channels.

This process is activated by the shared information of the bounding boxes on all channels, which realizes the highlighting of the target position information and plays a role of auxiliary activation. Besides, it highlights the location features of small targets, so that the network can better learn the feature information of small targets.
3. Experimental Results

3.1. Experimental Comparison and Analysis of Railway wagon Data Set

First of all, in order to verify that the network can better adapt to small targets, a high-definition line scan camera was used to collect photos of locks at the railway station. There are 1417 images in the training set, 50 images in the validation set, and 376 images in the test set. The categories include normal category and abnormal category.

The average input image size in this data set is 1380×1700, and the lock size is about 1/250 of the entire image area. Due to the large door, the proportion of the lock relative to the whole picture is too small, which belongs to the category of small target detection. Using K-Means dimensional clustering to generate 9 prior boxes. The initial learning rate is 0.001, adjusted to 0.0001 after 50 epochs. The convergence curve of the loss value during the training process is shown in Figure 3(a).

![Figure 3. Loss convergence curve.](image)

During the test, S-Darknet and today’s popular target detection models were used to train and detect the locks of railway freight trains in the same way, and compared their accuracy and speed. The specific comparison is shown in Table 1. It can be seen that the algorithm in this paper is superior to common target detection algorithms. In addition, the accuracy of the two categories in the data set was tested, and the P-R curve is shown in Figure 3(b).

![Figure 3. Loss convergence curve.](image)

**Table 1. Comparison of mAP value between S-Darknet and other algorithms.**

| Method    | FPS | mAP   | AP50  | AP75  |
|-----------|-----|-------|-------|-------|
| Faster RCNN | 8.5 | 86.59% | 88.32% | 87.10% |
The mAP of the S-Darknet network reached 92.86%, of which the accuracy rate of the abnormal type reached 87.42%, and the accuracy rate of the normal type reached 98.30%. Taking YOLOv4 as an example, Figure 4 is a test comparison diagram in the same picture. Experiments show that the S-Darknet network can avoid false detection of small targets. In the figure, the red box is the false detection, the pink box and the blue box are the labeled boxes, and the green is the detection box. It can be seen that the S-Darknet network can be detected correctly. Therefore, in the detection of very small objects, the algorithm in this paper has obtained better results.

Figure 4. YOLOv4 and S-Darknet detection comparison figure.

3.2. Comparison and analysis of PASCAL VOC data experiment

In order to verify that the S-Darknet network is not only suitable for the detection of small targets such as railway trucks, but also for the detection of small targets in the standard data set. Thus, this paper selects 20 categories in the VOC2012 and VOC2007 standard data sets for training and testing, and calculates different targets respectively. The mean average accuracy rate (mAP) of the detection algorithm on the data set. The test results are shown in Table 2. The mAP of the S-Darknet network can reach 74.2%. Compared with other classic target detection algorithms, it has greater advantages in terms of speed and accuracy. Compared with the recently proposed YOLOv4, the detection of small target categories and the overall mAP have a certain degree of improvement, and the speed is equivalent, and the effect of real-time detection can be achieved.

| Method       | S-Darknet | YOLOv3 | YOLOv4 | Faster R-CNN | SSD |
|--------------|-----------|--------|--------|--------------|-----|
| mAP/%        | 74.2      | 71.0   | 73.6   | 65.5         | 68.8|
| FPS          | 25.8      | 25.7   | 25.9   | 6.5          | 22.3|

The small target detection results of the S-Darknet network in the VOC2007 and VOC2012 data sets are shown in Figure 5. It can be seen that the algorithm can accurately detect small targets, and can also accurately detect when the background color is similar to the color of small objects.
4. Conclusion
This paper proposes the S-Darknet model, which is designed based on the Darknet-53 feature extraction network and can be used for the detection of railway freight locks. The effectiveness of the small target algorithm is proved by training and testing standard data sets. First, this paper designs a network framework with high integration, which can efficiently use the feature information of small targets of different scales, and it can also make full use of the semantic information of the context. Then, a feature enhancement module is proposed to enhance the target feature. The S-Darknet network has achieved good results through comparative experiments on the standard data set, and the accuracy of the actual data set has been significantly improved. It can be seen that this network can detect small targets well. In future research, we will consider improving the anchor matching strategy according to specific occasions to improve the generalization ability of the model.

Acknowledgments
Thanks for the help of classmates and teachers in the laboratory. Thanks to the National Natural Science Foundation of China (61170185, 61602320), the Liaoning Provincial Doctoral Startup Fund Project (201601172) and the Liaoning Provincial Department of Education's Scientific Research Project (L201607) for funding.

References
[1] KISANTAL, M, WOJNA, Z, MURAWSKI, J. (2019) Augmentation for small object detection. In: CVPR. Long Beach.1902.07296.
[2] Tang, X, Daniel, K.D., He, Z.Q., Liu, J.T. (2018) PyramidBox: A context-assisted single shot face detector. In ECCV. Munich.pp.797-813.
[3] Singh, B., Davis, L.S. (2018) An Analysis of Scale Invariance in Object Detection - SNIP. In: CVPR. Salt Lake City.pp.3578-3587.
[4] Girshick, R., Donahue, J., Darrell, T. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In:CVPR.Columbus.pp.580-587.
[5] Girshick, R. (2015) Fast R-CNN. In: ICCV. Santiago. pp. 1440-1448.
[6] Ren, S., He, K., Girshick, R., et al. (2015) Faster r-cnn: Towards realtime object detection with region proposal networks. In: NIPS. Montreal. pp. 91-99.
[7] Liu, W., Anguelov, D., Erhan, D., et al. (2016) SSD: single shot multibox detector. In: ECCV Amsterdam.pp.21-37.
[8] LIU, S., HUANG, D., WANG, Y., et al. (2018) Receptive Field Block Net for Accurate and Fast Object Detection. In: ECCV. Munich.pp.404-419.
[9] Redmon, J. et al. (2015) You Only Look Once: Unified, Real-Time Object Detection. In: CVPR. Boston.pp.779-788.
[10] Redmon, J. et al. (2017) YOLO9000: Better, Faster, Stronger. In CVPR. Honolulu.pp.6517-6525.
[11] Redmon, J. et al. (2018) YOLOv3: An Incremental Improvement. In: AVSS.Auckland.pp.1-6.
[12] Bochkovskiy, A., Wang C.Y. et al. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. In: CVPR.Seattle.2004.10934.
[13] Derakhshani, M.M., Masoudnia S., Shaker A.H. (2019) Assisted Excitation of Activations: A Learning Technique to Improve Object Detectors. In: CVPR. Seattle, pp. 9201–9210.