A Target Detection Algorithm based on Single Shot Detector

Yuanlin Feng*, and Inwhee Joe*
*Department of Computer Science, Hanynag University,
222 Wangsimni-ro, Seongdong-gu, Seoul, 04763, South Korea

Single Shot Detector 기반 타깃 검출 알고리즘

풍원림*,조인휘*
*한양대학교 컴퓨터소프트웨어학과
296786973@qq.com, iwjoe@hanyang.ac.kr

Abstract

In order to improve the accuracy of small target detection more effectively, this paper proposes an improved single shot detector (SSD) target detection and recognition method based on cspdarknet53, which introduces lightweight ECA attention mechanism and Feature Pyramid Network (FPN). First, the original SSD backbone network is replaced with cspdarknet53 to enhance the learning ability of the network. Then, a lightweight ECA attention mechanism is added to the basic convolution block to optimize the network. Finally, FPN is used to gradually fuse the multi-scale feature maps used for detection in the SSD from the deep to the shallow layers of the network to improve the positioning accuracy and classification accuracy of the network. Experiments show that the proposed target detection algorithm has better detection accuracy, and it improves the detection accuracy especially for small targets.

1. Introduction

Target detection plays an extremely important role in daily life [1-3], such as the field of drone aerial photography, remote sensing images, the detection of small and distant targets by unmanned vehicles, and the detection of small defects in products in the industrial field. and many more. Computer field [4-5] generally stipulates that objects with a size smaller than 32×32 pixels in an image are small targets. At present, the detection effect of large and medium targets in the target detection field has reached a very high level [6-7], but small targets have low resolution, blurred images, and less information, which leads to weak feature expression capabilities. Its detection has become a big problem in target detection.

The field of target detection has been developed for decades. Traditional target detection algorithms based on hand-designed features, such as Haar[8], HoG[9], etc., are gradually eliminated due to their low accuracy and poor robustness.

Following this is the development of various neural network algorithms based on deep learning that have been booming in recent years. These algorithms have greatly improved the detection accuracy while shortening the detection time. Detection algorithms based on convolutional neural networks can be roughly divided into two categories, one-stage (single-step) algorithm and two-stage (two-step) algorithm. The two-stage detection algorithm divides the detection task into two steps. The first step is to generate candidate regions, and the second step is to classify and predict the selected candidate regions. Although this method sacrifices speed, the detection accuracy is higher. The more representative ones are R-CNN[10] proposed by Girshick et al., and Faster-RCNN[11] proposed by Ren et al., which have achieved good detection results.

One-stage detection algorithm directly presets the size and aspect ratio of candidate frames in different regions, and then classifies and regresses, with fast speed and low accuracy. In 2016, Liu et al. proposed the single shot detector (SSD) [12] algorithm that uses features of different scales for feature extraction and fusion, which achieved the relative balance between accuracy and speed for the first time. It uses multi-scale features for small

Target detection, but because the receptive field of the low-level feature map used is not small enough, the SSD algorithm has a poor detection effect on small targets. In response to the above-mentioned problems encountered in target detection, this paper proposes an improved single shot detector (SSD target detection and recognition method) based on cspdarknet53, which introduces a lightweight ECA attention mechanism and a feature pyramid network (FPN).

First, replace the original SSD backbone network with cspdarknet53 to enhance the learning ability of the network.

Then, a lightweight ECA attention mechanism is added to the basic convolution block to optimize the network.

Finally, FPN is used to gradually integrate the multi-scale feature maps used for detection in the SSD from the deep to

the shallow layers of the network to improve the positioning accuracy and classification accuracy of the network. Experiments show that the proposed target detection algorithm has better detection accuracy, and the improved method has obvious effects, and has a certain detection accuracy improvement in target detection.

2. Proposed method

2.1 Single Shot Detector (SSD)

The main network structure of the SSD algorithm is VGG-16. The SSD network structure is shown in figure 1.

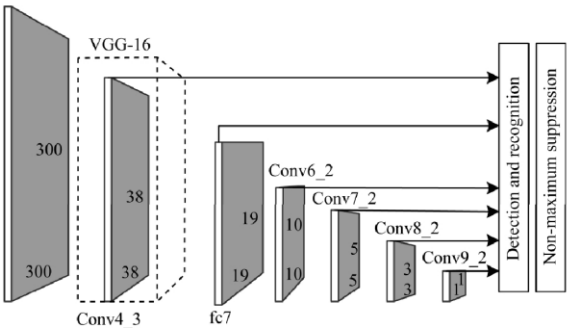

Fig. 1. The structure of SSD

The SSD algorithm obtains 6 different scale feature maps through the VGG network, the sizes of which are (38,38), (19,19), (10,10), (5,5), (3,3), (1 ,1), a series of fixed-size default boxes are set on the obtained feature maps, but the default box sizes corresponding to different feature maps are different. The calculation formula for selecting the default box on each feature map is:

$$S_K = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1), k = \{1, 2, \cdots, \ m\}$$

Among them, k represents the number of feature maps, Sm.im represents the ratio of the default box at the bottom layer to the input image, usually 0.2, and Sma, represents the ratio of the default box at the highest layer to the input image, usually 0.9. The width and height calculation formula of each default frame is as follows:

$$W_k = S_k\sqrt{r_n}, n = \{1, 2, \cdots, \ 5\}$$

$$\mathrm{H}_k = \frac{S_k}{\sqrt{r_n}}, n = \{1, 2, \cdots, \ 5\}$$

In particular, when r =1, the SSD algorithm also adds a default frame with a scale of 1 $S_{\mathrm{K}}^{'} = \sqrt{\mathrm{S_k S_{k+1}}}$, and the width and height $W_k^6 = H_k^6 = \sqrt{\mathrm{S_k S_{k+1}}}$ of the default frame are at this time. The comparison between the SSD default frame and the real frame under different scale feature maps is shown in figure 2.

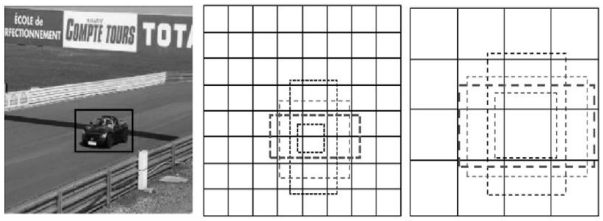

Fig. 2.different scale feature maps

2.2 Feature pyramid

Lin et al. [13] built a feature pyramid with marginal additional cost based on the inherent multi-scale hierarchical pyramid of deep convolutional neural networks, and proposed a deep-to-shallow architecture with lateral connections to build high-level semantic features at all scales. The graph is called a feature pyramid network, as shown in figure 3.

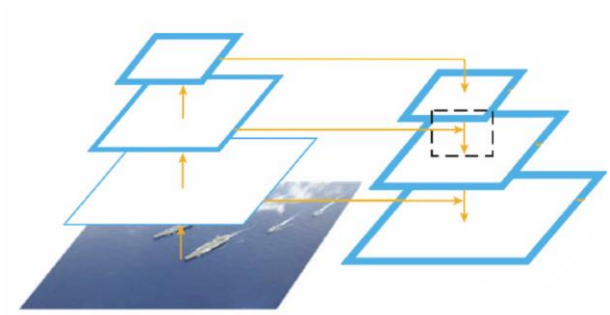

Fig. 3. Schematic diagram of feature pyramid network

For deep convolutional neural networks, the semantic information of the shallow features of the network is relatively small, but the location information is clear; the semantic information of the deep features of the network is rich, but the location information is rough. The feature pyramid network combines the semantic information of the deep network with the location information of the shallow network to improve the detection accuracy of the network.

The shallow-to-deep line is actually the forward process of the network. As the network deepens, the feature map will decrease from large to small, and the number of channels will increase to ensure the invariance of feature translation. The feature maps of the last few scales of the extracted network form a feature pyramid.

The deep-to-shallow circuit generally uses deconvolution or upsampling. Lateral connection (as shown in figure 4) is used to fuse the up-sampled feature map and the feature map generated by the 1×1 convolution kernel with the same width, height, and number of channels (corresponding elements are added). After fusion, a 3×3 convolution kernel will be used to convolve the fusion feature map, the purpose of which is to eliminate the aliasing effect of upsampling.

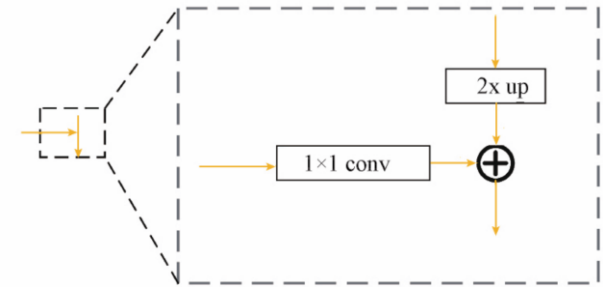

Fig. 4. Lateral connection

2.3 Efficient Channel Attention

The attention mechanism has been proven to improve the performance of CNN. SE-Net proposed an effective way to learn channel attention for the first time and achieved good performance. Subsequently, the development of attention module can be divided into two directions: (1) Feature fusion enhancement; (2) Channel and spatial attention combination. CBAM uses average and maximum pooling to fuse features. GSoP introduces second-order pooling to integrate features more effectively. GE uses deep convolution to explore spatial scalability and merge features. CBAM and scSE use a 2D convolution to calculate spatial attention. With a similar idea to Non-Local, GC-Net designed a simple NL network and integrated with the SE module to obtain a lightweight module to model long-term dependencies. Dual Attention Network (DAN) considers both NL-based channel attention and spatial attention in semantic segmentation. However, most NL-based attention modules can only be used in a single or a small number of convolution modules because of their high complexity. Obviously, all the above methods focus on how to design more complex attention modules to achieve higher performance. Unlike them, ECA is designed to learn effective channel attention with low complexity. The figure 5 shows the ECA module. After using GAP to aggregate the convolution features, there is no dimensionality reduction. The ECA module first adaptively judges the size of the convolution kernel, and then performs 1D convolution, followed by a Sigmoid function to learn channel attention. In order to apply ECA to CNN, the author replaced the SE module with an ECA module. This network is called ECA-Net.

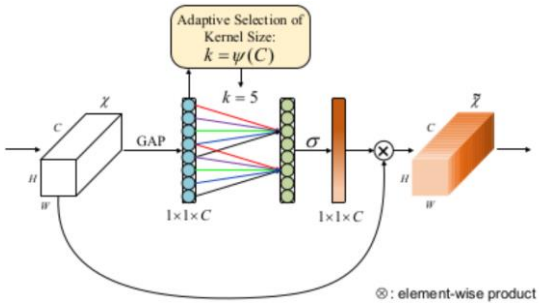

Fig. 5. The structure of ECA-Net

The ECA module intends to obtain local cross-channel communication, which is similar to channel local convolution and channel convolution; unlike them, ECA adjusts 1D convolution through adaptive convolution kernel size, replacing full connection in the channel attention module Floor. Compared with grouping and depth separable convolution, the method in this paper has better performance and lower complexity.

3. Experiment and Analysis

3.1 Data set and experimental environment

The experimental hardware adopts Z440 workstation with 32G memory and NVIDIA P2000 graphics card, and the operating system is Ubuntu 16.04. The programming environment used is as follows, the programming language used is Python language, and the deep learning framework is TensorFlow 2.0. The training samples for the experiment come from the training set in the Pascalvoc dataset.

3.2 Experimental Results and Analysis

In the training phase, each target in each image in the training set needs to be marked with a rectangular box, and the occluded target should also be marked. During the test, if the overlap between the identified target detection frame and the marked rectangular frame reaches more than 90% of the marked rectangular frame, the test is recorded as successful. The PASCAL VOC 2007 data set is a classic open source data set, including 5000 training set sample images and 5000 test sample images, and a total of 21 different object categories. This article uses the training set and test set of the data set, and the experiment uses the Tensorflow framework to implement the convolutional neural network model. The parameters such as random inactivation, maximum iteration value, batch size in SSD generate the average accuracy value (mAP) Greater impact. In order to get a better output, these parameters need to be optimized.

The detection result is shown in Figure 6, and the original SSD detection result is shown in the first row in Figure 6. The algorithm does not detect small objects in the image. The improved SSD network detection result is shown in the second row in Figure 6. The improved algorithm can detect small targets in the image.

Fig. 6. The result of Experiments.

The experimental results are shown in Figure 7. The blue bar graph represents the experimental results of the original SSD algorithm, and the red bar graph represents the experimental results of the improved SSD algorithm. The average AP before the improvement is 80.75%, and the average AP after the improvement is 83.92%. Comparing the experimental results, it can be concluded that the accuracy of the improved algorithm is increased by about 3.2%, indicating that the improved method is effective.

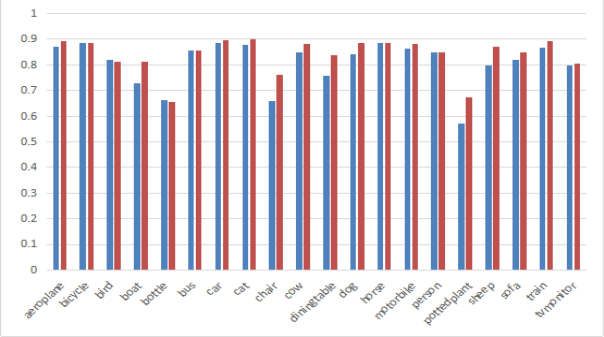

Fig.7. The accuracy of Experiments

Therefore, the improved model detection effect in this paper is relatively good. The experiments show that the proposed target detection algorithm has better detection accuracy, and the improved method has obvious effects. The detection accuracy of the target is improved to a certain extent, and the small target in the image can be detected.

4. Conclusion

Based on the problems of poor robustness and low positioning accuracy in traditional SSD algorithms, this paper proposes an improved single shot detector (SSD target detection and recognition method) based on cspdarknet53, which introduces lightweight ECA attention mechanism and Feature Pyramid Network (FPN). Firstly, replace the original SSD backbone network with cspdarknet53 to enhance the learning ability of the network. Then, a lightweight ECA attention mechanism is added to the basic convolution block to optimize the network. Finally, FPN is used to gradually integrate the multi-scale feature maps used for detection in the SSD from the deep to the shallow layers of the network to increase the receptive field of the low-level feature maps and improve the overall detection performance of the algorithm. The experimental results show that the improved SSD algorithm has better small target detection performance and greatly improves the algorithm's robustness.

References

[1] Fang LP, He H J, Zhou G M. Research overview of objectdetection methods[J]. Computer Engineering and Ap-plications, 2018, 54(13): 11-18

[2] Chen H J，Wang Q , Yang G w,et al. SSD target detection algorithm based on multi-scale convolution feature fu-sion [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 1049-1061.

[3] Huang J, Rathod v, Sun C, et al.Speed/accuracy trade-offs for modern convolutional object detectors[CV/ Proceedings of the 2017 IEEE Conference on Computer Vision andPattern Recognition，Honolulu，Jul 21-26, 2017. Pisca-taway: IEEE,2017: 3296-3297.

[4] Bell S, Zitnick C L, Bala K, et al.Inside-outside net: de-tecting objects in context with skip pooling and recurrentneural networks[J].IEEE Conference on Computer Visionand Pattern Recognition (CVPR), 2015: 2874- 2883.

[5] MaenpaaT,Pietikainen M. Texture analysis with local bi-nary patterns[J]. Handbook of Pattern Recognition & ComputerVision, 2005, 3540: 115-118.

[6] Dai J F, Li Y, He K M, et al. R-FCN: object detection viaregion-based fully convolutional net-works[CV/Proceedings of the 29th Annual Conference onNeural Information Processing Systems. Barcelona，Dec5- 10, 2016:379-387.

[7] Redmon J， Farhadi A. YOLO9000: Better，Faster，Stronger[CV/Proceedings of the 2017 IEEE Conference on Com-puter Vision and Pattern Recognition, Honolulu,Jul21-26,2017.Piscataway: IEEE, 2017: 6517-6525.

[8] Mita T, Kaneko T,Hori O. Joint Haar-like Features for Face Detection[J]. Proc.intl.conf.on Computer Vision, 2005, 2:1619-1626 Vol.2.

[9] Dalal N,Triggs B. Histograms of Oriented Gradients for Human Detection[CV/IEEE Computer Society Confer-enceon Computer Visi on and Pattern Recognition,San Di-ego,June20-26,2005.Piscata way:IEEE,2005:177-187.

[10] Girshick R B，Donahue J，Darrell T, et al. Regi on-based convoluti onal networks for accurate object detecti on andsegmentation[J]. IEEE Transactions on Pattern Analysi sand Machine Intelligence,2015,38(1): 142-158.

[11] Ren s Q,He K M，Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposalnetworks[C]//Proceedings of the 29th Annual Conferenceon Neural Information Processing Systems，Montreal. Dec7-12，2015. Red Hook: Curran Associates, 2015: 91 -99.

[12] Liu w, Anguelov D,Erhan D, et al. SSD: single shot multiboxdetector[C]V/LNCs 9905: Proceedings of the Eu-ropean Conference on Computer Visi on ，Amsterdam, Oct 11-14,2016.Berlin, Heidelberg: Springer, 2016: 21-37.

[13] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.