An Image Recognition Algorithm Based on Big Data Technology

Yijiang Ran*, Zejing Feng

College of Applied Science and Technology, Hainan University, Danzhou, Hainan 571737

*Corresponding author: yijiangran@hainanu.edu.cn

Abstract. In order to solve the problem of poor accuracy in small target recognition, a target detection algorithm guided by double attention model is proposed. Based on the implementation principle of the single-stage detection algorithm, this method introduces two attention models to improve the detection performance, especially for small target objects. Firstly, a multi-scale feature cascade attention module is introduced into the convolution neural network to pay attention to different regions in the feature map of the original convolution neural network, so as to reduce the background of the feature graph and the interference of negative sample information, especially in the shallow feature map, we can pay attention to small target objects effectively. In addition, the dense connection alleviates the problem of gradient disappearance in the process of network back propagation. Secondly, the significant channel self-attention module is introduced into the fused features to distinguish the different channels of the feature map and screen out the useful channel information to make the features to be detected more representative. The proposed method is tested on the target detection benchmark data set COCO to verify the effectiveness and advanced nature of the proposed method.

Keywords: Image processing, target detection, convolution neural network, small target, attention model.

1. Introduction

Target detection is one of the most important and challenging research topics in the field of computer vision. It requires the computer to classify the interested objects in an image containing multiple objects, and return the position of each object in the image through the boundary box. In practical application, target detection has very important research significance and value in target detection, precision guidance, intelligent monitoring, visual navigation, human-computer interaction, space remote sensing and medical assistant diagnosis.

In this paper, the idea of attention model is combined to improve the characteristics of convolution neural network, so as to improve the performance of the detector, especially the detection performance of small target objects. Considering the shortcomings of two-stage detection algorithm, such as large memory consumption and slow detection speed, a new single-stage detection algorithm called dual attention guidance model (DAGM) is proposed, which includes multi-scale feature cascade attention (MFCA) module and significant channel self-attention (SCSA) module.
2. Algorithm in this paper
The DAGM network structure model is shown in figure 1. Like most current single-stage detection algorithms, this framework adopts end-to-end training, which mainly includes feature extraction unit and feature prediction unit. In the feature extraction unit, MFCA module and SCSA module are added. Among them, the MFCA module produces an attention tensor in the process of feature extraction in each stage of the network, and carries on the weighted summation operation to each region of the feature map of this stage; the SCSA module distinguishes the importance of each channel of the feature map output of each stage of CNN. Through these two modules, the feature prediction unit is provided with features with stronger representation and richer semantic information. The feature prediction unit mainly includes the target classification module and the boundary box coordinate regression module. The target classification module uses four convolution layers, the size of the convolution kernel is $3 \times 3 \times 256$, the last layer uses the convolution kernel of $3 \times 3 \times d$, and the number of channels $d$ is $k \times n$, where $k$ is the number of types of the target object, $n$ is the number of anchor frames predicted on the feature map of each stage, and the nonlinear activation layer is used to calculate the prediction probability of each anchor frame. The regression module is roughly the same as the classification module, but the difference is that after the operation of four convolution layers, a convolution kernel of $3 \times 3 \times m$ is used in the last layer, and the number of channels $m$ is $4 \times n$.

![Figure 1 The structure of DAGM model in this paper](image1)

2.1. MFCA module
Following the idea of attention mechanism and hollow convolution, this method introduces MFCA module in the stage of CNN feature extraction to improve the feature expression ability of the network without excessively increasing the number of model parameters. The main operations include convolution block hollow block, up-sampling operation and cascade module (C). The specific connection mode of each component is shown in figure 2.

![Figure 2 MFCA module structure diagram](image2)

2.2. SCSA module
FPN structure fuses the high-level feature map of low-resolution and strong semantic information extracted by convolution neural network and the low-level feature map of high-resolution and weak semantic information in a top-down way, so that all levels of features contain rich semantic information and improve the accuracy of target detection, especially for small target objects.
Then the feature map obtained through the feature pyramid fusion operation does not pay attention to the different information between the channels, so a significant channel self-attention module is proposed. For different channels of the feature map, different weights are used to represent the importance of different channels, so that the feature map can be used more fully and effectively. The specific structure of the module is shown in figure 3.

![SCSA module schematic diagram](image)

The SCSA module first performs $3 \times 3$ convolution operation and nonlinear activation operation on the fused feature graph, and then carries out two pooling operations, including global maximum pooling (GMP) and global average pooling (GAP). Then the sum operation of the obtained vector is carried out, and finally, through the $1 \times 1$ convolution operation and activation operation, each channel of the input feature graph is given a meaningful weight. In order to reduce the amount of computation, only the significant channel attention operations are performed on the feature graphs with strong semantic information in the C3 and C4 and C5 stages of the ResNet-101 network, and different attention weights $W_k$: are assigned according to the differences among the channels of the input feature graphs.

$$W_k = \text{Sigmoid}(W_{\text{GAP}} + W_{\text{GMP}}),$$

$$P_k = S_k \cdot W_k$$

3. **Experimental results and analysis**

3.1. **Data set**
In order to verify the effectiveness of the proposed model, experiments are carried out on the COCO2017 benchmark data set. COCO dataset is an authoritative dataset currently used to evaluate the performance of target detection algorithms. The data set aims at the target detection task, including 80 kinds of target objects. The training set includes 118287 images to train the parameters of the model; the verification set includes 5000 images to verify the effectiveness of each module; and the test set includes about 20000 pictures to test the pros and cons of the model. However, performance indicators need to be tested on the official server. According to the pixel proportion of the target, the dataset divides the target into small targets ($S_{\text{area}} < 32 \times 32$, $S_{\text{area}}$ is the target area), medium targets ($32 \times 32 < S_{\text{area}} < 96 \times 96$) and large targets ($S_{\text{area}} > 96 \times 96$). The proportion of the number is 41%, 34% and 25%, respectively. According to the detection performance of these three kinds of size targets, the corresponding detection indexes APS, APM and APL are proposed. According to the different values of the cross-merge ratio (IoU), there are two main test indicators: AP50 and AP75, represent the average detection accuracy of all categories when the IoU value is 0.50 and 0.75, respectively. In addition, the most important index is that AP, averages the average detection accuracy of 80 target categories under 10 thresholds, which is the key to determine the quality of the detection algorithm.

3.2. **Experimental environment and training details**
The experiment is based on Keras deep learning platform and runs in Ubuntu14.04 system environment. The CPU is 3.3GHzIntelCore(TM)CPU i9-7900x, with 32GB memory, and the graphics card model is NVIDIA GeForce GTX1080ti, accelerator library, which is CUDA8.0 and CUDNN6.0.

The parameters of the backbone network are initialized on ImageNet. ResNet-101, is optimized by Adam optimizer. The exponential decay rate $\alpha$ is 0.9, $\beta$ is 0.999, and the learning rate is periodic learning...
rate. The initial learning rate is set to $5.0 \times 10^{-6}$, and the maximum learning rate is set to $1.5 \times 10^{-5}$. Due to the limited computing resources, batchsize is set to 1 in the experiment.

3.3. Contrast experiment

The popular target detection algorithms are selected and compared, including the two-stage detection algorithm FasterR-CNN and its extended algorithms, CoupleNet, SIN, DeNet and MLKP. There are also single-stage detection algorithms such as SSD513, YOLOv2, YOLOv3, DSSD513, RON and RelationNetwork. The experimental results are all obtained on the COCO test set, and the comparison results are shown in Table 1.

Table 1 Comparison of detection results of different detection algorithms

| Method                          | Backbone      | AP  | AP$_{50}$ | AP$_{75}$ | AP$_S$ | AP$_M$ | AP$_L$ |
|---------------------------------|---------------|-----|-----------|-----------|--------|--------|--------|
| Faster R-CNN + + + [8]          | ResNet-101-C4 | 34.9| 55.7      | 37.4      | 15.6   | 38.7   | 50.9   |
| Faster R-CNN (FR) by G-RMF [8] | ResNet-101-C4 | 34.7| 55.5      | 36.7      | 13.5   | 38.1   | 52.0   |
| YOLOv2 [16]                     | DarkNet-19    | 21.6| 44.0      | 19.2      | 5.0    | 22.4   | 35.5   |
| SSD513 [6]                      | ResNet-101-SSD| 31.2| 50.4      | 33.3      | 10.2   | 34.5   | 49.8   |
| DSSD513 [6]                     | ResNet-101-SSD| 33.2| 53.3      | 35.2      | 13.0   | 35.4   | 51.1   |
| RON [33]                        | VGG-16        | 27.4| 27.1      | 49.5      | —      | —      | —      |
| DeNet [31]                      | ResNet-101    | 33.8| 53.4      | 36.1      | 12.3   | 36.1   | 50.8   |
| CoupleNet [31]                  | ResNet-101    | 33.1| 53.5      | 35.4      | 11.6   | 36.3   | 50.1   |
| YoLoV3 [31]                     | DarkNet-53    | 33.0| 57.9      | 34.4      | 18.3   | 35.4   | 41.9   |
| SIN [31]                        | VGG-16        | 23.2| 44.5      | 22.0      | 7.3    | 24.5   | 36.3   |
| Relation Network [46]           | ResNet-50     | 32.5| 54.0      | 33.8      | —      | —      | —      |
| MLKP [336]                      | ResNet-101    | 26.9| 48.4      | 26.9      | 8.6    | 29.2   | 41.1   |
| MFCA (ours)                     | ResNet-101    | 36.2| 54.5      | 38.7      | 18.5   | 39.2   | 47.6   |

According to Table 1, the detection accuracy of most current detection algorithms for small target objects is low, and the detection accuracy of two-stage detection algorithms such as FasterR-CNN-based detection algorithm is higher than that of single-stage detection algorithm. While the proposed DAGM detection algorithm is better than the two-stage detection algorithm in terms of average detection accuracy, it also has the best performance for small target detection, which proves the effectiveness of the proposed algorithm.

3.4. Verification experiment of module validity

Experiments are carried out on the two proposed attention modules MFCA and SCSA to verify their effectiveness, and the test results are shown in Table 2.

Table 2 Comparison of test results of each module

| Module                     | Backbone      | AP  | AP$_{50}$ | AP$_{75}$ | AP$_S$ | AP$_M$ | AP$_L$ |
|----------------------------|---------------|-----|-----------|-----------|--------|--------|--------|
| Baseline                   | ResNet-101    | 35.0| 52.6      | 37.7      | 17.1   | 38.8   | 48.0   |
| Baseline + MFCA            | ResNet-101    | 35.5| 53.3      | 38.1      | 18.4   | 39.2   | 48.1   |
| Baseline + MFCA + SCSA     | ResNet-101    | 35.9| 54.0      | 38.6      | 18.8   | 39.8   | 48.5   |

Table 2 the experimental results are all tested on the COCO verification set. Compared with the benchmark structure, after adding the MFCA module, the AP increased by 0.5%, and the APS for small target objects increased by 1.3%, and the detection performance improved significantly; then after adding the SCSA module on the basis of the MFCA module, the AP increased by 0.4%, and the detection accuracy of large, medium and small target objects was improved. The coordination of the two modules designed in the experiment shows that compared with the benchmark recognition framework, the AP, AP50 and AP75 of the proposed algorithm model DAGM on COCO data sets are increased by 0.9%,
1.4% and 0.9%, respectively. This is consistent with the motivation of this paper, using effective attention in each region and channel of the feature to highlight the information of the target object in the feature graph, which is equivalent to a screening of the channel of the feature graph, and the top-down feature fusion enriches the context information of the feature graph, which is more conducive to the detection of target objects, especially the detection performance of small target objects is significantly improved. The effectiveness of these two modules is proved.

4. Conclusion
A target detection algorithm based on DAGM is proposed. The algorithm includes two kinds of attention modules. One is that in the process of feature extraction by convolution neural network, the MFCA module is designed to give different importance weights to each region of the feature graph. Second, the SCSA module is introduced into the fused feature map, which makes the differences between different channels of the feature map, and realizes the adaptive learning of features, which strengthens the information of the target object and effectively removes the information interference of background and negative samples in the image. The effectiveness of each module is explored, and the experimental results show that the proposed algorithm has better detection performance.

References
[1] Tan Guanghong, Hou Jin, Han Yanpeng, et al. Low-parameter real-time image segmentation algorithm based on convolution neural network [J]. Progress in Laser and Optoelectronics, 2019556 (9): 091003.
[2] Liu Yuzhen, Jiang Zheng, Ma Fei, et al. Hyperspectral image classification based on hypergraph and convolutional neural network [J]. Advances in Laser and Optoelectronics, 20119, 56 (11): 111007.
[3] Liu Fang, Wang Xin, Lu Lixia, et al. Geomorphological image classification based on sparse coding and convolutional neural network [J]. Journal of Optics, 20119, 39 (4): 041000.
[4] Xu Yan, Sun Meishuang. Convolution neural network image defog algorithm based on multi-feature fusion [J].