Pedestrian detection based on attention mechanism and feature enhancement with SSD

T T Feng
School of Information Science and Technology
Donghua University
Shanghai, China
1497243382@qq.com

H Y Ge
School of Information Science and Technology
Donghua University
Shanghai, China
gehuayong@dhu.edu.cn

***Abstract*—Some factors such as low resolution of small targets, limited target features, and noise interference affect the effect of pedestrian detection. Therefore, a pedestrian detection algorithm based on attention mechanism and feature enhancement with SSD is presented in this paper. It uses channel feature fusion to fuse non-adjacent convolutional layers to obtain significant edge gradient features and semantic information features.Finally, through the optimization of the attention mechanism CBAM, the channel features and spatial features are coupled under different fusion detection layers to improve the feature weight of the pedestrian's salient region, so as to the detection accuracy of the algorithm has taken a big step forward. The improved algorithm is verified on the fusion dataset and the VOC2007TEST dataset, compared with the SSD algorithm, the detection accuracy of the improved algorithm model on the fusion dataset reaches 60.1%, with an increase of 7.2%, and that on the VOC2007TEST dataset reaches 78%, with an increase of 4.1%. The experimental results show that this method can effectively detect the small target pedestrian in the image, and reduce the error detection and omission detection, thus verifying its feasibility.**

Keywords—SSD, CBAM, pedestrian detection, feature enhancement

I. INTRODUCTION

An important branch of the computer vision field is pedestrian detection, and its technology has been applied to more fields.There are two types of pedestrian detection algorithm based on deep learning, such as Fast R-CNN[1]and Faster R-CNN[2-3], which are target detection algorithms based on candidate regions with tedious training steps, slow speed and occupying too much physical space, while regression detection algorithms are represented by YOLO[4] and SSD[5].YOLO algorithm divides feature maps into n×n grids, but sometimes there are many small targets in a grid, which is likely to cause missed and false detection of targets,and SSD algorithm can extract multiple candidate regions in a grid, taking into account both the accuracy and speed of detection, but its ability to express the characteristics of the network is still insufficient, such as poor robustness, poor boundary positioning and other problems. Inspired by DSSD[6] algorithm, a deconvolution layer is added after SSD auxiliary convolutional layer to form an "hourglass" structure of "wide-narrow-wide", which has a greater improvement in small target detection. In [7] and [8], the channel attention mechanism SENet is used to screen out the information that needs to be retained, which improves the ability to detect small targets,so the pedestrian detection algorithm presented in this paper picks the feature enhancement method in the framework of the SSD algorithm to fuse non-adjacent feature layers to increase the semantic information and contour gradient information of the detection layer, and add the attention mechanism module CBAM[9] after the fusion of the feature layers, it retains more target feature information and suppresses irrelevant information, which greatly improves the pedestrian detection accuracy of the SSD algorithm.

II. SSD ALGORITHM FRAMEWORK AND PRINCIPLE

SSD adopts VGG16 as the basic model, and uses convolutional layer to replace two fully connected layers, namely FC6 and FC7, while increasing the depth of convolutional layer, namely four convolutional layers of Conv8_2, Conv9_2, Conv10_2 and Conv11_2. Its model structure is shown in Fig.1.

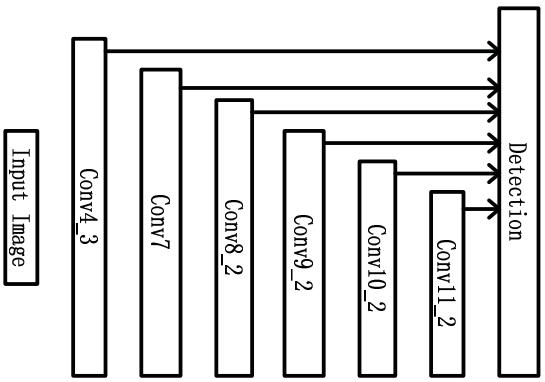


Fig.1. SSD model structure diagram.

SSD discretizes the output space of each detection layer into a series of default boxes, namely, candidate boxes, which are finally summarized together for the NMS algorithm to effectively deal with multi-scale problems.

978-1-7281-8589-7/20/$31.00 ©2020 IEEE

III. SSD PEDESTRIAN DETECTION ALGORITHM COMBILING ATTENTION MECHANISM AND FEATURE ENHANCEMENT

The feature layer of SSD algorithm is in the form of pyramid, the lack of the characteristic information of each other between different scales,and one-eighth of the feature information of the lowest level detection layer Conv4_3 is lost, while the sensing field of the feature layer is large at the higher level, and its feature extraction shows that the data content is too abstract.These problems have limitations in the time of dealing with small target pedestrian with SSD.

The attention mechanism in deep learning obtains the target area that needs to be focused on by quickly scanning the global text, and then obtains the detailed information by investing more computing resources in the focused target area. This mechanism greatly improves the efficiency and accuracy of visual information processing.

In summary, this paper will combine the attention mechanism CBAM and feature enhancement strategy to improve the SSD model.

A. Attention mechanism CBAM

Attention mechanism CBAM is a lightweight general module that saves parameters and computing power. For a given feature map, the weight of attention is deduced in turn along the two dimensions of space and channel, and then adjust the features adaptively by multiplying with the original feature map. The realization of attention mechanism consists of two modules: channel attention and spatial attention.

Each channel of the feature represents a special detector, so it makes sense for the channel attention to focus on what kind of feature.The structure of the channel attention module [9] is shown in Fig.2.

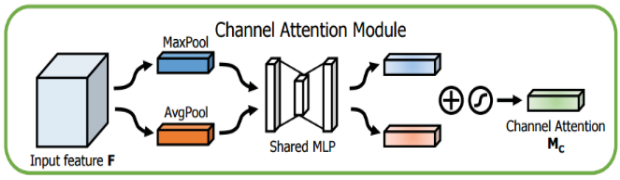


Fig.2. Schematic diagram of channel attention module.

Its calculation formula is as follows:

$$\begin{aligned} \boldsymbol{M}_c(\boldsymbol{F}) &= \sigma(\boldsymbol{MLP}(\boldsymbol{AvgPool}(\boldsymbol{F})) + \boldsymbol{MLP}(\boldsymbol{MaxPool}(\boldsymbol{F}))) \\ &= \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{avg}^c)) + \boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{max}^c))) \end{aligned} \quad (1)$$

The specific operation steps are: enter a feature F of $\boldsymbol{H}\times\boldsymbol{W}\times\boldsymbol{C}$, where H, W and C represent the length, width and number of channels of the feature map. First, the spatial information of the feature graph is aggregated in two ways: global average pooling AvgPool and maximum pooling MaxPool, which are based on length and width, and two different $1\times1\times\boldsymbol{C}$ channel spatial feature $\boldsymbol{F}_{avg}^c$ and $\boldsymbol{F}_{max}^c$ are generated, indicating the average pooling features and maximum pooling characteristics.Then, they are respectively passed through a shared two-layer neural network MLP. To reduce the parameters, the number of neurons in the first layer is C/R, r represents the compression ratio, the activation function is Relu, and that in the second layer is C. Then, add the two features together and get the weight coefficient through a Sigmoid activation function. Finally, the new scaled feature $\boldsymbol{M}_c(\boldsymbol{F})$, namely the input feature required by the spatial attention module, can be obtained by multiplying the weight coefficient and the original feature F. In (1): σ represents Sigmod activation function; the weights of the shared fully connected layer MLP are $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ respectively.

The channel attention module is then connected to the spatial attention module to focus on the feature information of the target's location, that is, where the feature is meaningful, which is a supplement to the channel attention feature map. To calculate spatial attention feature maps, two operations of average pooling and maximum pooling are used along the channel, and they are connected to generate valid features. In the information region, the application of pooling can effectively improve the significance of the target characteristics in the channel. The structure of the spatial attention module [9] is shown in Fig.3.

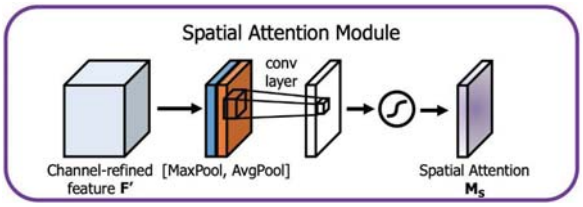


Fig.3. Schematic diagram of spatial attention module.

Its calculation formula is as follows:

$$\begin{aligned} \boldsymbol{M}_s(\boldsymbol{F}) &= \sigma\left(\boldsymbol{f}^{7\times7}\left(\left[\boldsymbol{AvgPool}(\boldsymbol{F});\boldsymbol{MaxPool}(\boldsymbol{F})\right]\right)\right) \\ &= \sigma\left(\boldsymbol{f}^{7\times7}\left(\left[\boldsymbol{F}_{avg}^s;\boldsymbol{F}_{max}^s\right]\right)\right) \end{aligned} \quad (2)$$

The specific operation steps are:given a $\boldsymbol{H}\times\boldsymbol{W}\times\boldsymbol{C}$ feature $\boldsymbol{F}'$,average pooling and maximum pooling of channel dimensions are carried out respectively to obtain two $\boldsymbol{H}\times\boldsymbol{W}\times1$ channel space features $\boldsymbol{F}_{avg}^s$ and $\boldsymbol{F}_{max}^s$, and the two channel space features are splintered together along the channel dimension. Then, through a 7×7 convolutional layer, the activation function is Sigmoid, and the weight coefficient is obtained. Finally, the new scaled feature is obtained by multiplying the weight coefficient and the feature F'.In (2), $\boldsymbol{f}^{7\times7}$ is the convolution operation, the shape of the convolution kernel is 7×7.

Typically combining the two modules of channel attention and spatial attention in a sequential manner, and putting channel attention in front can achieve better results.The CBAM attention mechanism module [9] structure is shown in Fig.4.

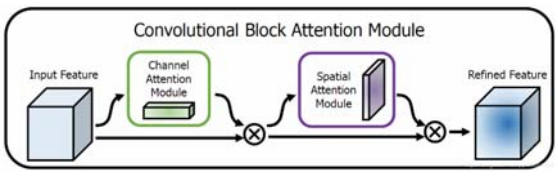


Fig. 4. Schematic diagram of CBAM attention mechanism module.

B. SSD model of embedding feature enhancement strategy

A new algorithm, feature enhancement, is proposed in this paper. For the sake of using a certain method to generate new features from the extracted different features and make the new features more effective for classification, three high-low level network fusion modules are set up,that is,Conv11_2, Conv10_2, Conv9_2 are up-sampled and concatenated with Conv4_3, Conv7, and Conv8_2 respectively.Then perform feature fusion of channel and space different weighting coefficients on the fused feature layer, and finally classification and regression are performed with the multi-scale characteristic layers inherent in SSD.Therefore, the final model proposed in this paper is shown in Fig.5.

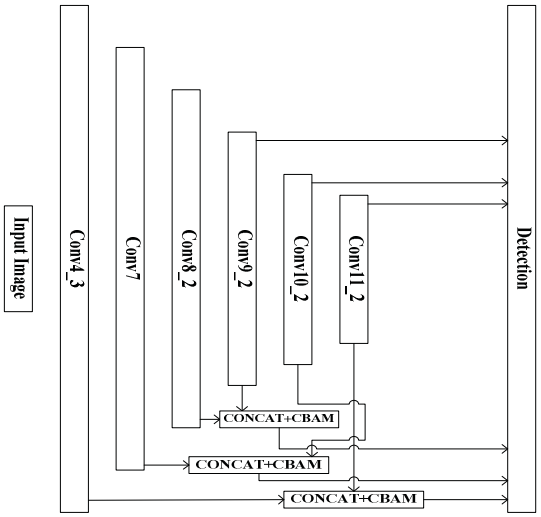


Fig. 5. Framework of our whole module.

In Fig.5, "CONCAT+CBAM" refers to the mutual splicing between feature layers, after which the CBAM module is added. Although the CBAM module is embedded after the fused feature layer, what hasn't changed is the shape and size of the feature map. Thence, the shape and size of each detection layer of the improved model of the SSD are the same as before.

IV. EXPERIMENT AND RESULT ANALYSIS

A. Experimental environment and dataset

The configuration used in this experiment is the Ubuntu 16.04 operating system, the GPU of GTX1080Ti is selected, the deep learning framework is Tensorflow, and the development language is Python 3.5. In this paper, the fusion dataset is used for training, including VOC2007, VOC2012, COCO, SYSU, and PRW dataset, then the fusion dataset and VOC2007TEST dataset are used for testing. Among them, there are 3776 fusion datasets in the test set.,the SYSU dataset includes pedestrians in parks and subways, the PRW dataset is some pedestrian pictures taken continuously by the camera in the Sun Yat-sen University.In VOC2007, VOC2012, and COCO, only some photos of pedestrians are extracted; in VOC2007TEST, only 2097 photos of pedestrians are extracted.

B. Experimental results and comparative analysis

This experiment uses AP as the evaluation index of the model, taking into account both precision and recall, and is often used as an evaluation index for multi-target detection models.In this experiment, the batch size is set to 32, the initial learning rate is 0.001, the first 500 steps of the learning rate adopt the “warm up” strategy, and the remaining steps adopt the piecewise constant attenuation method, and iterate 120,000 steps to obtain the final network model. This paper compares the proposed SSD pedestrian detection algorithm based on the attention mechanism and feature enhancement with the SSD. The PR curve and AP value of the fusion dataset and the VOC2007TEST dataset containing only pedestrians are shown in Fig.6 and Fig.7.

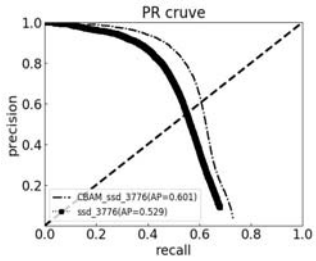


Fig.6.Test results of fusion dataset.

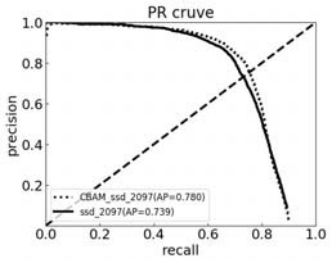


Fig.7. Test results of VOC2007TEST dataset.

The PR curve shows that the proposed method is superior to all other methods at most thresholds.With the help of feature enhancement information, each detection layer contains more contour information and stronger semantic information, and the edge information of the calculation result is clear. At the same time, the designed attention mechanism makes the experiment positioning accuracy high, so as to get a better PR curve.

By combining the attention mechanism and feature enhancement strategy to optimize the SSD algorithm, with the input of resolution, the improved algorithm compared with the original SSD, the fusion dataset AP value increased from 0.529 to 0.601, an increase of 1.4 percentage point, the AP value of the VOC2007TEST dataset increased from 0.739 to 0.780, an increase of 1.2 percentage points.These results veryfy the utility of the attention mechanism and feature enhancement strategy, and improve the detection effect of small target pedestrians.

The algorithm in this paper is also compared with similar pedestrian detection algorithms. Comparing the methods of [10] and [11] with the algorithm in this paper, the test results are shown in tab.1.

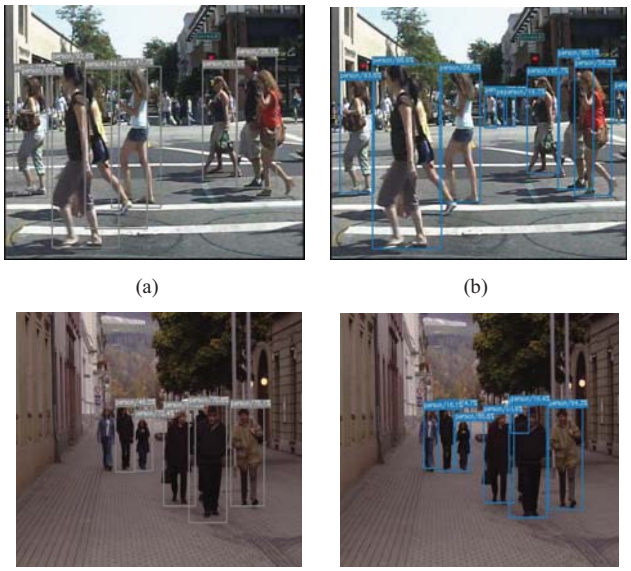
TABLE I. COMPARISON OF RESULTS IN DIFFERENT METHODS.

methods		***MAP/%***	
		VOC2007TEST	fusion dataset
Reference [10] original method		49.31	
Reference [10] improved method		60.76	
Reference [11] original method		72.3	
Reference [11] improved method		74.2	
this paper	SSD	73.9	52.9
	SSD improvement	78.0	60.1

Literature[10] added residual connection structure and Batch Normalization on the basis of SSD network, the AP

value on VOC2007TEST dataset increased from 0.493 to 0.608, an increase of 11.45 percentage points; literature[11] replaced the traditional NMS algorithm with Soft-NMS algorithm on the basis of Faster Rcnn network, and strengthened the ability of Faster R-CNN algorithm to identify overlapping areas. At the same time, the algorithm replaces the uniformly sampled anchor points with "Hot Anchors", and the AP value on the VOC2007TEST dataset increases from 0.723 to 0.742, an increase of 1.9 percentage points.Compared with the improved methods in [10] and [11], the improved algorithm in this paper has increased by 17.24% and 3.8%, respectively. The reason is that attention mechanism CBAM and feature enhancement strategy are added in this paper, which can quickly screen out high-value information from a large amount of information with limited attention resources, so as to improve the efficiency of the neural network.

The comparison between the visual inspection results of the improved SSD algorithm and the SSD algorithm on the fusion dataset is shown in Fig.8.It was found in the experiment that there are two main factors that affect the detection effect, which are wrong detection and missed detection. Among them, the left side (a) and (c) of Fig.8 are the SSD detection results, and the right side (b) and (d) of Fig.8 are the detection results of the improved SSD algorithm. Comparing figure (a) and figure (b), we can see that the algorithm in this paper detects four more targets than SSD, and one target is missed; comparing figures (c) and (d) show that the algorithm in this paper detects two more targets than the SSD, and the SSD misses one target and the disadvantage is that the algorithm in this paper only frames the upper half of a pedestrian, but not the entire pedestrian. According to the experimental results, generally speaking, the improved SSD algorithms have better detection effects.



(a) (b)

(c) (d)

Fig. 8. Comparison of SSD algorithm before and after improvement.

V. CONCLUSION

This paper introduces the feature enhancement strategy and attention mechanism module CBAM into the popular one-stage target detection framework SSD, and proposes SSD pedestrian detection based on the attention mechanism and feature enhancement.On the one hand, the feature enhancement strategy can effectively enhance the feature information of each other between the SSD feature layers. On the other hand, the attention mechanism CBAM can optimize the channel and spatial features and enhance the detection accuracy of small targets. By verifying the fusion dataset and the VOC2007TEST dataset, we compared the algorithm in this paper with other popular pedestrian detection algorithms. It is found through experiments that the algorithm in this paper has certain recognition ability for small targets, which improves the problems of missed detection and poor positioning of targets, and greatly improves the robustness of the algorithm. How to further simplify the network structure, reduce model parameters and calculations, achieve embedded use or further improve the detection accuracy and reduce the false detection rate and missed detection rate is the direction of the next effort.

ACKNOWLEDGMENTS

As this paper is about to be completed, I would like to thank my mentor--associate professor HuaYong Thanks to my family for their infinite care and love. I wish them good health and everything is well! Finally, I would like to express my sincere thanks to the teachers who have worked so hard to review this article!

REFERENCES

[1] Girshick R, “Fast R-CNN," International Conference on Computer Vision ,2015,pp. 1440-1448.

[2] Ren S, He K, Girshick R and Sun J,“ Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,pp. 1137-1149.

[3] Sun X, Wu P and Hoi S C, “Face Detection using Deep Learning: An Improved Faster RCNN Approach," Neurocomputing, 2018, pp. 42-50.

[4] Redmon J, Divvala S K, Girshick R and Farhadi A, “ You Only Look Once: Unified, Real-Time Object Detection,"Computer Vision and Pattern Recognition,2016, pp. 779-788.

[5] Wei L, Dragomir A, Dumitru E, Christian S, Scott R, Cheng Yang F and Alexander C Berg,“SSD: Single Shot MultiBox Detector," 2015.

[6] Cheng Y F,Wei L, Ananth R, Ambrith T and Alexander C Berg, “ DSSD : Deconvolutional Single Shot Detector," Computer Vision and Pattern Recognition,2017.

[7] Mo R J, Jiang N L, Zhong B W and Hai B L, “ Multi-scale target detection algorithm based on attention mechanism,"Journal of Optics,2020, pp. 1-15.

[8] Hai T Z and Meng Z, “SSD detection algorithm based on attention mechanism, " Computer Engineering,2020, pp. 1-8.

[9] Woo S, Park J, Lee J Y and Kweon I S , CBAM: Convolutional Block Attention Module (Springer, Cham vol 11211) ed Ferrari V, Hebert M, Sminchisescu C and Weiss Y .Switzerland: Computer Vision ,2018, pp. 3-19.

[10] Tao W, Cui P S, Li J W, Yao Z, Yuan P and Qi H P, “ Research on pedestrian detection method based on SSD target detection algorithm," Scientific and Technological Innovation,2020, pp. 62-63.

[11] Wan Y Y and Jin P L,“Pedestrian detection algorithm based on improved Faster R-CNN," Science, Technology and Engineering,2020, pp. 1498-1503.