Yolov5-DP: A New Method for Detecting Pedestrian Aggregation

Kunzhi Yang[1], **Xiaoyan Chen**[1*], **Xiaoning Yan**[2], **Dashuo Wu**[2]

[1]*Tianjin University of Science and Technology, China*[*]

[2]*Shenzhen Softsz Co. Ltd., China*

E-mail: 1663491981@qq.com

www.tust.edu.cn

Abstract

In this paper, a novel network Yolov5-DP (Yolov5-DBSCAN-P) is proposed. Deep separable convolution and ACON-C activation function are added into Yolov5 network to improve the detection accuracy of pedestrians. Firstly, DBSCAN-P is used as the clustering detector to detect pedestrians in the area. Secondly, the depth-separable convolution is used to replace the common convolution in Yolov5. Finally, the loss function Swish is improved to ACON to increase the model speed and reduce the model size. The Yolov5-DP network is tested on the public dataset MOT20Det. The experimental results show that good detection results and accurate aggregation detection results are obtained.

Keywords: Target detection, YOLOv5, clustering, regional population analysis

1. Introduction

Both at home and abroad, there are tragedies caused by crowds gathering every year. However, in daily life, schools, factories and other places often show the phenomenon of crowd gathering; During rallies, crowds will inevitably gather in squares and streets. In the face of this crowd load, every extreme behavior is very easy to cause accidents. In this context, in order to reduce the occurrence of this kind of tragedy, this paper aims to study the pedestrian detection task in (gathering) area.

In recent years, most research on pedestrian detection at home and abroad is based on object detection in computer vision. The development of target detection has roughly experienced two periods: traditional target detection (before 2014) and target detection based on deep learning (after 2014)[1]. The traditional target detector has sliding window (VJ), HOG, DPM, etc. With the manual characteristic performance becoming saturated, it is replaced by deep learning. Nowadays, target detectors based on deep learning have become common, which can be divided into Two categories: two-stage Detection based on candidate regions and one-stage Detection based on regression classification. Two-stage target detector is represented by RCNN[2] series, while one-stage target detector is represented by SSD[3], RetinaNet[4], YOLO[5] series.

The main research direction of crowd is crowd counting, and the key technology is to integrate the generated thermal map to get crowd counting. There are also crowd-based abnormal (clustering) behavior research, which has many methods, including traditional methods based on machine learning, big data analysis, and methods combined with deep learning and machine learning, with few data.

2. Yolov5-DP Network

YOLOv5 model is a recently proposed neural network model, which performs very well in target detection tasks. Excellent results have been achieved in both COCO and VOC datasets. YOLOv5 network structure is shown in Fig.1.

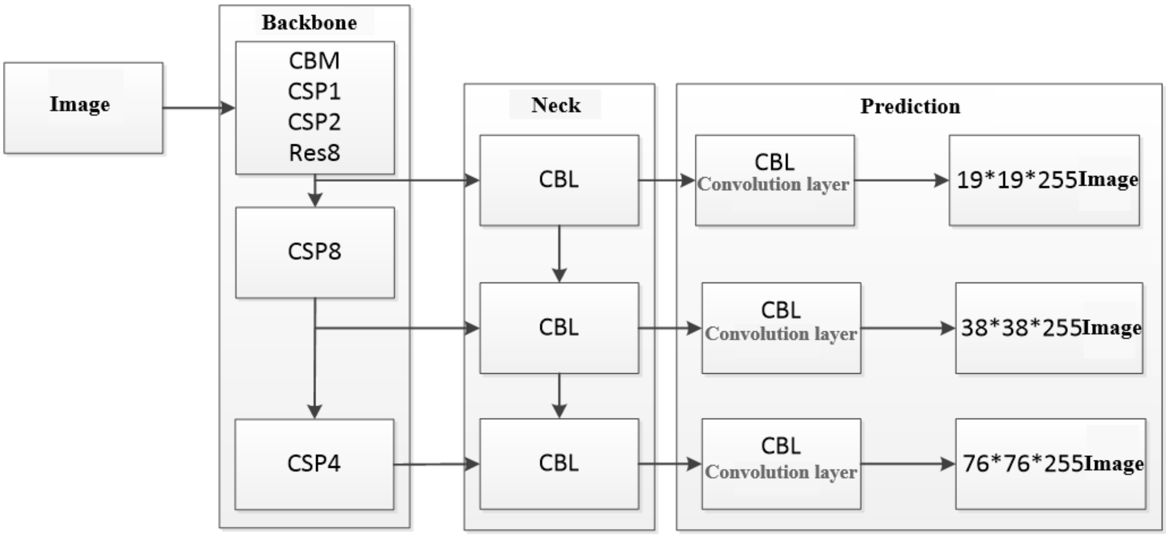

Fig. 1. YOLOv5 network structure

1.1. *Depthwise separable convolution*

Depthwise Separable Convolution has two steps: Depthwise Convolution and Pointwise Convolution. First, each feature graph at each layer corresponds to a convolution with a convolution kernel size of 3×3 for convolution operation, and then carries out the second convolution operation with a convolution kernel size of 1×1 for each channel to obtain the output feature graph with the set size.

Let the size of depthwise convolution kernel be $D_K * D_K * 1$, the number of input channels C_{in}, the number of output channels C_{out}, W and H are the width and height of the input feature graph respectively. The formula for calculating the proportion of parameter number (param) and computation amount (FLOPs) between depth-separable convolution and ordinary convolution is as follows:

$$\eta_{param} = \frac{D_K * D_K * C_{in} + C_{in} * C_{out}}{D_K * D_K * C_{in} * C_{out}} = \frac{1}{N} + \frac{1}{D_K^2} \qquad (1)$$

$$\eta_{FLOPs} = \frac{D_K * D_K * C_{in} * W * H + C_{in} * C_{out} * W * H}{D_K * D_K * C_{in} * C_{out} * W * H} = \frac{1}{N} + \frac{1}{D_K^2} \qquad (2)$$

In this paper, deep separable convolution is used to replace ordinary convolution in YOLOv5 networks. The advantage of using depth separable convolution is that the number of parameters and computation are smaller than ordinary convolution. It can be seen from the formula that the number of parameters and calculation quantity are $\frac{1}{N} + \frac{1}{D_K^2}$ of ordinary convolution, and the value is between $1/8$ and $1/9$ when the convolution kernel size is 3×3.

1.2. *ACON-c activation function*

Ma N et al proposed a new activation function ACON (Activate or not)[6]. ACON has several variants, and

ACON-C is one of the most widespread forms. ACON-c simply uses hyperparameters to scale the feature graph, and the formula is:

$$f_{ACON-C}(x) = (p_1 - p_2)x \cdot \sigma[\beta(p_1 - p_2)x] + p_2 x \qquad (3)$$

Unlike the first derivative of Swish's activation function[7], which has fixed upper and lower bounds, the first derivative of ACON-C has learnable upper and lower bounds.

2. Algorithm Testing and Experimental Analysis

2.1. *Selection and processing of experimental data*

In the pedestrian detection algorithm based on YOLOv5 network, the selection of data sets has a great influence on the detection effect, and there are certain requirements on the total amount of data images or the number of single objects used for training model. MOT20Det open source dataset was selected for the dataset. The visualization of the MOT20Det dataset for different scenarios is shown in Fig.2.

The amount of data used for train and VAL (verification) in the training were 8000 groups and 931 groups respectively, and the grouping was randomly selected. There were 4479 sets of data used for tests.

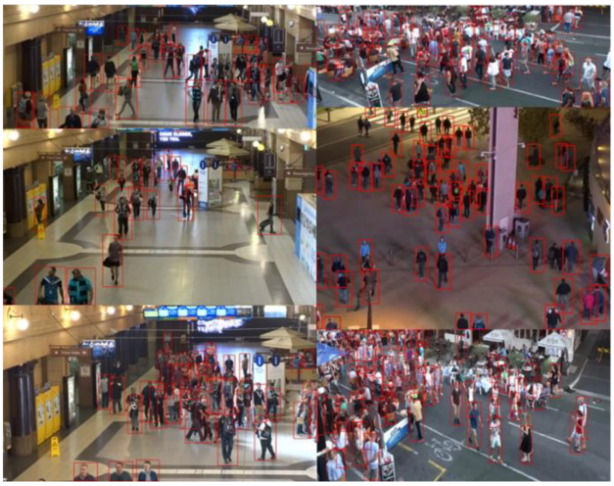

Fig. 2. Visualizations of different scenarios from the MOT20Det dataset

2.2. *Pedestrian detection experiment and test*

3.2.1.*Model training*

As the task only needs to detect pedestrians, the smallest model YOLOv5s is selected to ensure high accuracy and high performance pedestrian detector with high detection efficiency.

The convolutional substitution depth detachable convolution experiment in YOLOv5 model, this paper designed three detachable convolution structures for comparative experiments, as shown in Fig.3, Fig.4 and Fig.5.

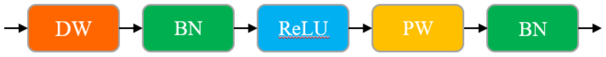

Fig. 3. Structure of DP 1

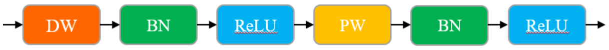

Fig. 4. Structure of DP 2

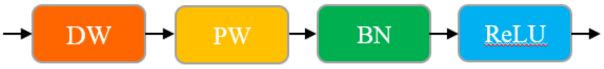

Fig. 5. Structure of DP 3

3.2.2.*Experimental results and analysis*

In Table 1, Conv represents the original YOLOv5 model, DP_3_S is to replace the ReLU activation function in DP_3 with SiLU, DP_3_ac is the replacement of ACON-C activation function, and DP_3_F is the depth-separable convolution replacement except the Focus layer. The activation function of YOLOv5 model from input layer to SPP layer is replaced by SiLU with ACON-C, that is, ACON-C.

YOLOv5 has 1×1 convolution and 3×3 convolution. According to the experiment, compared with only replacing 3×3 convolution, the model accuracy of replacing all convolution remained unchanged, and the number of parameters and training time increased a little. Therefore, the subsequent deeply separable convolution replacement experiment only targeted at 3×3 convolution.

It can be seen from Table 1 that the single ACON-C activation function replacement has a small increase in accuracy at the cost of a larger model. DP_3_S structure has the best performance in the deeply-detachable convolution substitution experiment. The combination of ACON-C function and deeply-detachable convolution can accelerate the network training speed on the premise of keeping the accuracy unchanged. Compared with the original YOLOv5, the size of the DP_3_ac model is reduced by half while the accuracy is basically the same.

Another conclusion is that the training speed is not proportional to the size of the model.

Table 1. Comparison of YOLOv5 model adjustment results

Method	mAP@0.5	mAP@0.5:0.95	param	GFLOPs	epochs	duration/h
Conv	0.9673	0.7904	7063542	16.4	100	5.932
ACON-C	0.968	0.7922	7408118	18.1	100	6.296
DP_1	0.9559	0.7359	3194746	6.7	100	8.205
DP_2	0.9582	0.745	3194746	6.7	100	8.17
DP_3	0.96	0.7558	3190114	6.6	100	12.387
DP_2_S	0.9641	0.7721	3194746	6.7	100	7.772
DP_3_S	0.9649	0.7785	3190114	6.6	100	7.191
DP_3_F_S	0.9645	0.7767	3193078	7.2	100	7.801
DP_3_ac	0.9648	0.7793	3534690	8.3	100	6.418

2.3. ***Pedestrian Cluster Detection Test and Testing***

3.3.1. *Algorithm implementation details*

In the experiment of clustering detection algorithm, there are data processing and parameter selection. Each part is introduced as follows:

(1) Firstly, the normalized coordinate data of pedestrian detection frame is extracted, and the pixel coordinates of the center point are generated into array data.

(2) K-means algorithm ADAPTS to elbow method to get the optimal clustering result; The MinPts parameter of DBSCAN algorithm is set to 3, and eps value is obtained by the adaptive algorithm. The MinPts parameter in the OPTICS algorithm is 5 (default); HDBSCAN algorithm has no manual parameter input. The MinPts parameter of DBSCAN -p is set to 3, and the proportionality coefficient n is set to 1.25.

3.3.2. *Performance evaluation index*

Since there is no performance evaluation index for clustering detection, this paper carries out clustering detection based on clustering algorithm, and the performance measurement of clustering is selected as one of the performance evaluation indexes for clustering detection. Real-time detection is also particularly important, so the time complexity of the algorithm is also used as the comparison item of the aggregation detection experiment.

The performance measures of clustering are divided into internal indicators and external indicators, both of which are effective performance indicators of clustering. The external indicators need to be evaluated according to the reference model. The MOT20Det data set used in the experiment has no aggregation data annotation. Therefore, this paper only compares the internal indicators to judge the rationality of clustering within the cluster, and then makes a comparative analysis on the clustering effect, that is, the visual images of clustering detection. Silhouette Coefficient, a common internal indicator, is adopted in this paper [8]

3.3.3. *Experimental results and analysis*

The 139 groups of data used in this experiment were randomly extracted from all 13410 groups of data in MOT20Det dataset at a ratio of 1%, including 8 groups of lens framing images of training set and detection set. The effect map in the experiment only visualized the aggregation target with rectangular boxes of different colors. Experimental data of aggregation detection algorithm comparison are shown in Table 2, where the effective data amount is the reference sub-data of contour coefficient s(i).

Table 2. Performance comparison of clustering detection algorithms

methods	Silhouette Coefficient/s(i)	Valid data quantity/sheet	Time complexity/s
K-Means	0.4301	139	0.13519
DBSCAN	0.3221	129	0.00135
OPTICS	0.5814	124	0.04315
HDBSCAN	0.4502	103	0.00257
DBSCAN-P	0.6269	63	0.05815

Different from k-means algorithm, the clustering result must be greater than or equal to 2 clusters, DBSCAN series algorithms will produce clustering results with the number of clusters equal to or less than 1. However, both contour coefficient method and DI index and DBI index, both common internal indicators, need to be used in the case of more than one cluster, so the experiment only carries out statistics on valid data. The contour coefficient s(i) value in Table 2 is the average sum of s(i) of each group of effective data, and the time complexity is the average sum of operation time of each algorithm on 139 groups of data.

3. Algorithm testing and experimental analysis

In this paper, the method of target detection is used to analyze and study pedestrian detection in the region. For pedestrian detection, this paper adjusted the YOLOv5 network and conducted a comparative experiment on the original model and the model with depth detachable volume and ACon-C activation function added to each structure combination, specifically for the comparison of training and test results. Finally, the optimized pedestrian detector is obtained. For clustering detection, four classical clustering algorithms including K-means, DBSCAN, OPTICS and HDBSCAN and the dbSCAN-P clustering algorithm adjusted based on DBSCAN in this paper were compared and tested to obtain clustering detector. The optimized YOLOv5 was used as the pedestrian target detector, and dbSCAN-P algorithm was used for clustering detection. Finally, the two were verified on the MOT20Det test dataset to complete the study of pedestrian detection in the area based on YOLOv5.

References

1. Zou, Zhengxia, et al. "Object Detection in 20 Years: A Survey." ArXiv Preprint ArXiv:1905.05055, 2019.
2. Girshick, Ross, et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
3. Liu, Wei, et al. "SSD: Single Shot MultiBox Detector." 14th European Conference on Computer Vision, ECCV 2016, 2016, pp. 21–37.
4. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318–327.
5. Redmon, Joseph, et al. "You Only Look Once: Unified, Real-Time Object Detection." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
6. Ma N , X Zhang, Sun J . Activate or Not: Learning Customized Activation. 2020.
7. Ramachandran, Prajit, et al. "Searching for Activation Functions." ArXiv: Neural and Evolutionary Computing, 2017.
8. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Authors Introduction

Mr. Kunzhi Yang

He received his bachelor's degree in Engineering from Lanzhou University of Technology in 2019. He is currently studying control science and engineering at Tianjin University of Science and Technology.

Prof. Dr. Xiaoyan Chen

She is currently a professor and doctoral supervisor at the School of Electronic Information and Automation, Tianjin University of Science and Technology. Her research interests are pattern recognition and deep learning.

Mr. Xiaoning Yan

He received his Master's degree in engineering from the University of Texas at Dallas in 2015. He is currently the Technical director of Shenzhen Ansoft Huishi Technology Co., LTD.

Mr. Dashuo Wu

He received his master's degree in engineering from Tianjin University of Science and Technology in 2019. He is currently an algorithm engineer of Ansoft Technology, engaged in deep learning and computer vision algorithm research, mainly involving target detection research direction.