A package auto-counting model based on tailored YOLO and DeepSort techniques

Sijun Xie[1,2], *Yipeng* Zhou[1, 2,*], *Iker* Zhong[2, 3], *Wenjing* Yan[1,2], and *Qingchuan* Zhang[1,2]

[1]School of E-business and Logistics, Beijing Technology and Business University, Beijing, 100048, China
[2]National Engineering Laboratory for Agri-Product Quality Traceability, Beijing, 100048, China
[3]School of Computer Science, Beijing Technology and Business University, Beijing, 100048, China

Abstract. In the industrial area, the deployment of deep learning models in object detection and tracking are normally too large, also, it requires appropriate trade-offs between speed and accuracy. In this paper, we present a compressed object identification model called Tailored-YOLO (T-YOLO), and builds a lighter deep neural network construction based on the T-YOLO and DeepSort. The model greatly reduces the number of parameters by tailoring the two layers of Conv and BottleneckCSP. We verify the construction by realizing the package counting during the input-output warehouse process. The theoretical analysis and experimental results show that the mean average precision (mAP) is 99.50%, the recognition accuracy of the model is 95.88%, the counting accuracy is 99.80%, and the recall is 99.15%. Compared with the YOLOv5 combined DeepSort model, the proposed optimization method ensures the accuracy of packages recognition and counting and reduces the model parameters by 11MB.

Keywords: Object tracking, Object detection YOLOv5, DeepSort, compressed, Deep learning model.

1 Introduction

With the development of computer vision, object recognition and tracking methods based on deep learning are becoming more and more popular. Among them, the YOLO (you only look once) has attracted much attention since it was proposed in 2016. The YOLO is a deep learning model in object detection with an outstanding performance in speed and precision. Combined with the tracking method—DeepSort, the construction is widely applied in industry, agriculture, and transportation area [1-4]. In order to obtain higher accuracy, the general trend of object recognition and tracking is to make more in-depth and complex networks [5-8]. However, advances in accuracy do not necessarily make the recognition more efficient in terms of scale and speed. In many industrial applications, such as automatic

* Corresponding author: yipengzhou@163.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

transmission, object counting, and video surveillance, recognition tasks need to be performed in a timely manner on a computing constrained platform.

This paper aims to study lighter deep learning construction of object recognition and tracking base on tailored YOLO (T-YOLO) and DeepSort methods. Also, it purposes a counting strategy to realizing the package counting in the input-output process of the warehouse. The experimental results show that the T-YOLO method can ensure recognition accuracy as well as greatly reduces the model parameters. In practical application, the model is more suitable for the deployment of the industrial terminal and it realizes the automatic counting task of the package in and out of a warehouse.

2 YOLOv5 architecture

The YOLOv5 framework mainly consists of three components, including backbone network, neck network, and detect network [9-11]. In the YOLOv5 model, it designs two Cross Stage Partial (CSP) structures, where CSP1 is used in the backbone network, and CSP2 is used in the neck network. On the backbone network, CSP net structure and focus structure are mainly used. In the CSP of the backbone network, CBL is composed of Convolution(Conv) layer, BatchNormalization(BN) layer, and Leaky ReLU. The key step of focus structure is slicing, which is shown in Figure 1. The slice operation changes the original 416 * 416 * 3 images into 208 * 208 * 12 feature images and then performs a Convolution operation with 32 Convolution cores to turn it into 208 * 208 * 32 feature images. The main purpose is to minimize the loss of information.

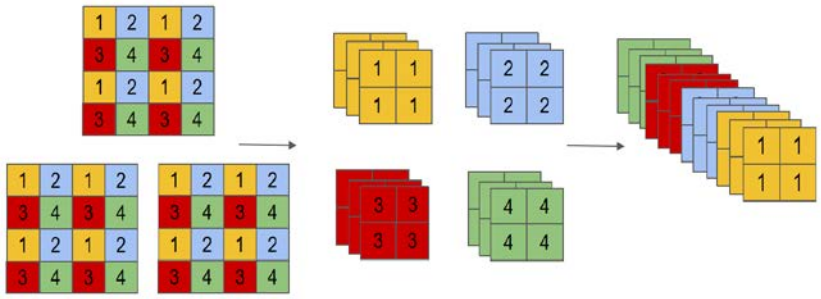

Fig. 1. Schematic diagram of images slice.

In the neck network, Feature Pyramid Networks (FPN) structure and Path Aggregation Network(PAN) structure were used on Neck. FPN is top-down, which uses the way of upsampling to transfer, and fuse information to obtain the predicted feature maps. PAN structure is bottom-up, and the information is transferred and fused by downsampling.

The detect network is mainly used for the final detection part of the model, which applies anchor boxes on the feature map output from the previous layer, and outputs a vector with the category probability of the target object, the object score, and the position of the bounding box surrounding the object [9].

3 T-YOLO combined DeepSort construction

3.1 Structure of T-YOLO

In this study, we cut two structures in the backbone, one is Conv, the other is BottleneckCSP. After cutting these two structures, the number of convolution cores generated in the backbone network is reduced, and the width of the network is also reduced accordingly. Due

to the cutting of the structure, the corresponding Concat structure in the neck network has also changed. Therefore, the size of the three prediction feature maps generated by T-YOLO is consistent with that generated by the original network structure, but the number of Convolution cores is reduced, which is in line with the warehouse package (such as unlabeled sugar packages) with simple features and middle target size.

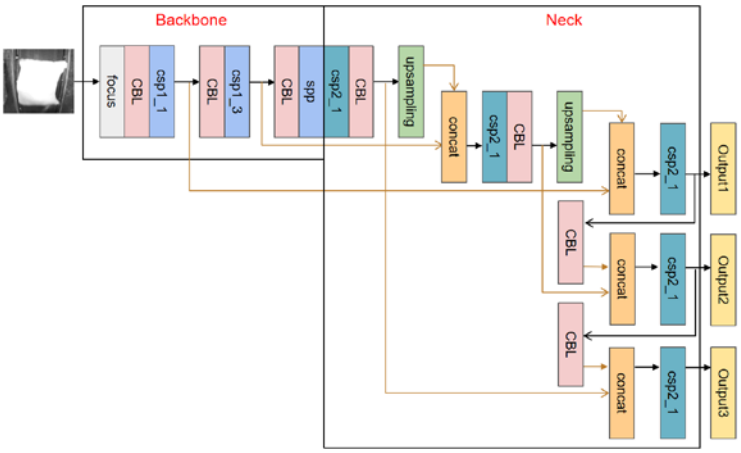

Fig. 2. Structure of T-YOLO.

3.2 DIoU_NMS

In the post-processing process of target detection, Distance Intersection over Union (DIoU) is applied in Non-Maximum Suppression (NMS) operation to select bounding boxes. The DIoU-NMS is defined as:

$$si = \begin{cases} si, & IoU - R_{DIoU}(M, Bi) < \varepsilon, \\ 0, & IoU - R_{DIoU}(M, Bi) \geq \varepsilon, \end{cases} \tag{1}$$

where M is the high confidence candidate box, and Bi is the coincidence of the traversed boxes and M. In addition, R_{DIoU} is the distance between the center points of two boxes, which is expressed by the following formula:

$$R_{DIoU} = \frac{\rho^2(b, b^{gt})}{c^2}, \tag{2}$$

where band b^{gt}represent the center points of the anchor frame and target frame respectively, and ρ represents the Euclidean distance between the two center points, and c represents the diagonal distance of the smallest rectangle that can cover the anchor and the target box at the same time.

3.3 DeepSort

DeepSort uses $(u, v, \lambda, h, x', y', r', h')$ to describe the state of the trajectory at a certain time. The u, v is the center coordinates of the bounding box, λ is the aspect ratio, and h is the height. The other variables are the respective velocities of the variables. Then a Kalman filter is used to predict the updated trajectory, which adopts a uniform velocity model and linear observation model to predict u, v, λ, h. Mahalanobis distance is used to describe motion matching degree between positions, which is defined in the following formula:

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i), \quad (3)$$

where d_j is the position of the j-th detection frame, y_i is the predicted position of the i-th tracker to the target, S_i is the covariance matrix between the detection position and the average tracking position. If the Mahalanobis distance of an association is less than the specified threshold t, the association successfully sets the motion state.

$$b_{i,j}^{(1)} = 1[d^{(1)}(i,j) \leq t^{(1)}], \quad (4)$$

3.4 Counting Strategy

In this study, three-fourths of the length of the screen area which is placed in the middle is selected as the recognition center to ensure that the package can be identified in a stable scene. Then, the DeepSort algorithm cascades the packages whose feature matching degree is greater than 0.7 and counts the packages when 8 consecutive frames are matched successfully.

4 Experimental and results

4.1 Experimental data set

First of all, the image data obtained by dividing the monitoring video into frames. In many pictures, there are no packages on the conveyor belt, which belong to the background pictures, there is no practical significance for this study, so this part of the picture will be removed. After that, the LabelImg tool is used to label the images. In manual operation, only the user-defined label needs to be marked in the image, and the tool can automatically generate the corresponding configuration file. Finally, after filtering out the unlabeled data, 1800 images are generated as the data set of training.

4.2 Experimental results

The training loss curve of the T-YOLO model and YOLOv5 model is shown in Figure 3. It can be seen that the training loss curve of the T-YOLO model is consistent with that of the YOLOv5 model. In the first 50 periods of network training, the loss value of the two models has decreased rapidly, and basically, tends to be stable after 200 epochs of training. Therefore, the model output after 300 epochs of training was determined as the package target recognition model in this study.

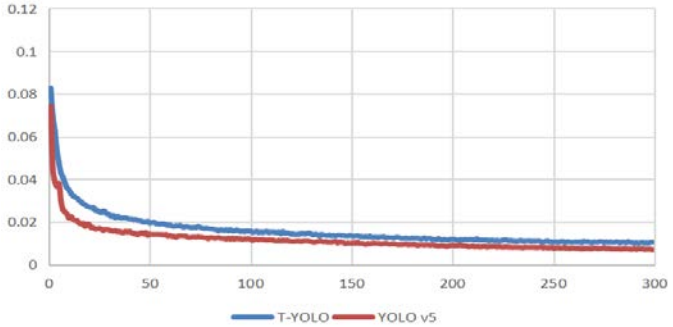

Fig. 3. Loss function value of two models.

In order to verify the performance of the T-YOLO model in warehouse package recognition, this experiment trains the data set for 300 iterations and tests the model. Figure 4a shows the result of single package identification, and Figure 4b shows the result of overlapping package identification. It can be seen that the T-YOLO model can correctly identify packages in two different cases, and has certain robustness.

(a) (b)

Fig. 4. Package recognition. (a) single packet recognition (b) overlapped packets recognition.

The comparison of experimental results between the YOLOv5 combined DeepSort construction and the T-YOLO combined DeepSort construction in the warehouse packages target detection are shown in Table 1. From the experimental results, it can be seen that the packet recognition accuracy of the YOLOv5 combined DeepSort construction is 96.23%, the mAP is 99.73%, the recall is 100%, and the counting accuracy is 99.23%. The packet recognition accuracy of the T-YOLO combined DeepSort construction is 95.88%, the mAP is 99.50%, the recall is 99.15%, and the counting accuracy is 99.80%. Although the recognition accuracy of the T-YOLO combined DeepSort construction is 0.35% lower than that of the YOLOv5 combined DeepSort construction, the parameter file is reduced by nearly 11MB. Therefore, for the identification task of the warehouse packages, in the case of a large number of packages, the T-YOLO combined DeepSort construction takes into account the recognition accuracy and parameter optimization, and can better complete the task of package identification.

Table 1. Experimental results of warehouse package on two construction.

Construction	Recognition accuracy/%	Recall/%	mAP/%	Parameter/MB	Counting accuracy/%
YOLOv5 combined DeepSort	96.23	1	99.73	14.4	99.23
T-YOLO combined DeepSort	95.88	99.15	99.50	3.8	99.80

5 Conclusion

The T-YOLO combined DeepSort target recognition and tracking construction proposed in this paper can greatly reduce the parameters while maintaining the recognition accuracy. It is suitable for real-time object detection and tracking, such as package counting in the input-output warehouse process.

References

1. Li H, Deng L, Yang C, et al. Enhanced YOLOv3 Tiny Network for Real-Time Ship Detection From Visual Image[J]. IEEE Access, 2021, 9: 16692-16706.
2. Wang X, Liu J. Tomato Anomalies Detection in Greenhouse Scenarios Based on YOLO-Dense[J]. Frontiers in Plant Science, 2021, 12: 533.
3. Laroca R, Zanlorensi L A, Gonçalves G R, et al. An efficient and layout- independent automatic license plate recognition system based on the YOLO detector [J]. IET Intelligent Transport Systems, 2021, 15(4): 483-503.
4. Knausgård K M, Wiklund A, Sørdalen T K, et al. Temperate fish detection and classification: A deep learning based approach[J]. Applied Intelligence, 2021: 1-14.
5. Han B G, Lee J G, Lim K T, et al. Design of a Scalable and Fast YOLO for Edge-Computing Devices[J]. Sensors, 2020, 20(23): 6779.
6. Gao J, Chen Y, Wei Y, et al. Detection of Specific Building in Remote Sensing Images Using a Novel YOLO-S-CIOU Model. Case: Gas Station Identification[J]. Sensors, 2021, 21(4): 1375.
7. Sun Z, Huang L, Jia R. Coal and Gangue Separating Robot System Based on Computer Vision[J]. Sensors, 2021, 21(4): 1349.
8. Hu X, Liu Y, Zhao Z, et al. Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network[J]. Computers and Electronics in Agriculture, 2021, 185: 106135.
9. Yan B, Fan P, Lei X, Liu Z, Yang F. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. Remote Sensing. 2021; 13(9):1619.
10. Schmidhuber J (2015) Deep learning in neural networks: anoverview. Neur Netw 61:85–117
11. White D, Svellingen C, Strachan N (2006) Automated measure-ment of species and length of fish by computer vision. Fish Res80(2-3):203–210