ISSN: 2664-9640

Research on Package Detection Algorithm Based on Convolutional Neural Network

Jie Jian [a], Shiming Yang [b], QiWei Chen [c]

[1] School School of Economics and Management, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[a]jianjie@cqupt.edu.cn, [b]2018211011@stu.cqupt.edu.cn, [c]565335478@qq.com

Abstract

This article aims to solve the actual logistics transfer center scene.This paper establishes a package inspection data set based on the target detection data set Pascal VOC data set and MS COCO data set to provide a data basis for solving the actual problem of package inspection. At the same time, the package detection algorithm is improved to make the prediction frame deviate from the surrounding non-corresponding target frame, which greatly improves the detection accuracy. Given the problem of small target packages in package detection, the appropriate anchor size is determined, and combined with the FPN feature fusion method, the detection effect of small target packages is significantly improved, and the detection accuracy is greatly improved.

Keywords

Deep learning, package detection, object detection, convolutional neural network, feature fusion.

1. Introduction

With the rapid development of e-commerce and the Internet, logistics and people's lives are getting closer and closer. According to data released in January 2020, China's express delivery business volume and business revenue were 63 billion pieces and 745 billion yuan respectively in 2019, an increase of 24% and 23% year-on-year. It is a very difficult task to deliver so many express deliveries to customers safely and in good condition. Taking the logistics transit center as an example, it is inevitable that express parcels will jam and fall during the transportation process, which reduces the operational efficiency of the transit center and increases the management cost of package transportation. In this context, the concept of smart logistics came into being. Smart logistics (Liu et al., 2020; Sedaghat et al., 2011) refers to the realization of automated transportation and efficient management of goods through information technologies such as big data and the Internet of Things.

Package inspection is one of the basic tasks for monitoring violent violations of packages, identifying jams, and falling abnormal packages. The purpose of package inspection is to determine whether there is a package in the image or video, and if it exists, return the check box representing the location of the package. Through the detection and positioning of the package on the scene monitoring video data of the logistics transfer center, the monitoring system can timely understand the current express delivery status, and make timely judgments on abnormal conditions such as jams and falling of the package, thereby effectively improving the logistics transfer center Operational efficiency reduces the manpower cost of management. Nowadays, the work of parcel management is mainly done manually, and the staff with a low level of automation use surveillance video to look for dropped or lost parcels. Not only does this require a lot of time and energy, but manual inspections are prone to fatigued workers, omitting some important information, and increasing the error rate. In actual scenarios, the problem of

ISSN: 2664-9640

parcel loss often occurs, which requires staff to pass surveillance video inspections, which takes a lot of time. Therefore, the rapid and accurate detection of packages in the logistics transfer center has great application prospects and practical significance.

2. Research on Package Detection Based on Repulsion Function

2.1. Package detected data set

2.1.1. Package detected data set establishment

Training package based on the detection depth study of the package requires a large amount of data during the test, the data needs to be labeled. This article uses labeling open-source software to label data, specify the area where the package is located in the image, and draw the bounding box.

2.1.2. Evaluation index

This article will use the same model evaluation criteria as Pascal VOC and MS COCO. That is, when the IoU(Intersection over Union)of the prediction box and the GT(Ground Truth)is greater than a certain threshold, it is regarded as a correct prediction. The accuracy of the network model will be calculated by calculating the AP (Average Precision) to evaluate.

Pascal VOC contains 20 types of objects to be detected. All prediction boxes with IoU greater than 0.5 are regarded as correct predictions. The AP values of each type are counted, and all types of AP are averaged to obtain mAP (mean Average Precision), reflecting the overall network Detection accuracy.

2.2. Design of package detection network based on Faster R-CNN

2.2.1. Feature extraction network

The VGG network has excellent feature extraction capabilities, strong generalization capabilities, excellent effects in semantic segmentation, target detection, and other fields, and has a wide range of applications. VGG-16 is divided into 16 layers, composed of 13 convolutional layers and 3 fully connected layers. All of them use 3×3 convolution kernels, which are simple to implement. This article chooses the VGG-16 network to extract feature maps and uses them in The VGG-16 weights trained on ImageNet are used as the initial weights of the feature extraction network, and the final feature map will be used for the classification and positioning of the package target.

2.2.2. Area generation network

The convolutional neural network can effectively extract the features in the image. In the target detection, the convolutional neural network can be directly used as the feature extraction network to obtain the key information in the image. Selective Search generates a series of candidate frames through a heuristic algorithm, inputs the image area framed by the candidate frames into the convolutional neural network, obtains its feature expression, and then uses the classifier to classify the detection results.

This method faces serious efficiency problems, (Ren et al., 2017)proposed Faster R-CNN, One of the most important points is to propose an RPN（Region Proposal Networks）. RPN is an algorithm for generating candidate frames in Faster R-CNN. RPN replaces the original sliding window and SS (Selective Search) method of generating candidate frames, and directly makes a sliding window in the feature map of the last layer of the convolutional neural network to generate candidate frames, reducing convolution operations and improving candidate frames The speed of generation is one step closer to real-time target detection.

ISSN: 2664-9640

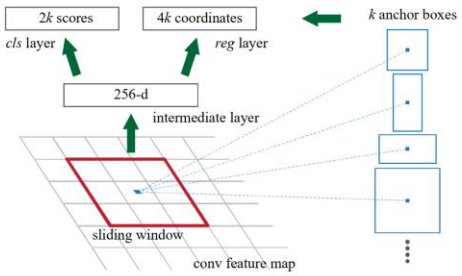

Figure 1 RPN structure diagram

The RPN network input can be an image of any size and outputs a series of rectangular target boxes, each of which is accompanied by a classification score. To generate candidate regions, a small size n×n window is slid on the feature map output by the last convolutional layer. Each sliding window gets a low-dimensional feature vector through mapping. This feature vector then inputs to two parallel fully connected layers for positioning and classification respectively. Since this window slides on the entire image, all the parameters of the fully connected layer are for the entire image and are global. The anchor introduced in Faster R-CNN describes multiple prediction candidate areas corresponding to a single sliding window, that is, a single sliding window corresponds to multiple anchors, and these anchors may become the final output candidate area.

The RPN network only classifies anchors into two categories, namely target, and background. The positive sample can be considered as the anchor with the highest overlap rate with the label frame, or the anchor with the overlap rate exceeding 0.7, so one label frame may correspond to multiple positive samples. The RPN network continues to use the multi-task comprehensive loss in Fast R-CNN and uses the loss function of classification and position regression to minimize the objective function.

2.2.3. Target classification and localization

(1) ROI pooling

After the RPN network obtains the feature map, to be unified into a fixed size, a fixed size feature map is obtained from a deep network with multiple convolution kernel pooling.

First, according to the input image, map the ROI to the corresponding position of the feature map, and then divide the mapped area into sections of the same size (the number of sections is the same as the output dimension), and finally, each section is max pooling Obtain fixed-size corresponding feature maps from boxes of different sizes, which is beneficial to input to the next layer of the network, and improves the network's processing speed of feature maps.

(2)Bounding box regression

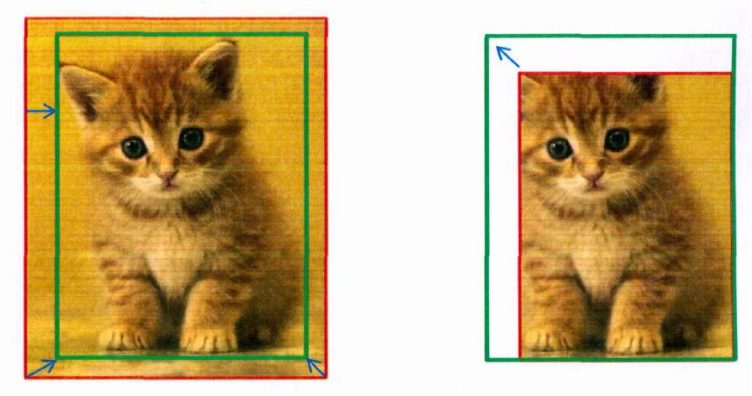

Figure 2 The change of detection frame

ISSN: 2664-9640

For each subregion of the image, in addition to classification operations, you can also fine-tune the range of the region. As shown in Figure 4, when the detection frame (red frame) is too large, the content in the frame can be judged to appropriately tighten the detection frame to a more suitable position (green frame); when the detection frame is too small, it can also be based on the image Content to guess the approximate location of the green box. In short, when the detection frame is close to the outer frame of the object, this fine-tuning of the detection frame can be tried.

In the current object detection framework, the outer border regression has become the default component (Girshick et al., 2014; Felzenszwalb et al., 2010; Hoiem, 2012; Erhan et al., 2014).In the outer frame regression, use the regression method to adjust the position of the detection frame (including size and aspect ratio).To make the regression target independent of the initial position and reduce the difficulty of the regression problem, the residual of the outer frame is returned, that is, the offset from the initial frame (red frame) to the target frame (green frame) in the figure. To make the regression goal meet the additivity, the coordinates of the box are parameterized:

$$t_x = (x_t - x_s)/w_s \tag{2-1}$$

$$t_y = (y_t - y_s)/h_s \tag{2-2}$$

$$t_w = log(w_t/w_s) \tag{2-3}$$

$$t_h = log(h_t/h_s) \tag{2-4}$$

x, y, w, h are the x Direction center y Direction center, width, and length, x_t, x_s corresponding to the target frame (green frame) and initial frame (red frame), (t_x, t_y, t_w, t_h) is the four-dimensional vector as the regression target.

In the training phase, the algorithm takes the sub-image in the detection frame as input,(t_x, t_y, t_w, t_h) is the regression target training regressor (regression tree, support vector regression (SVR, neural network, etc.)).In the test phase, the regression device obtains the output according to the input image content; the output is de-parameterized to fine-tune the detection frame.

Bounding box regression plays a key role in improving the accuracy and speed of object detection algorithms. From the perspective of accuracy, the outer frame regression can improve the alignment performance of the frame in the three dimensions of space, scale, and aspect ratio. The sliding window detection framework is limited by the computational cost. Only a few points can be discretely sampled for detection in the two dimensions of the frame's scale and aspect ratio. The fine-tuning ability of the outer frame regression helps to compensate for the performance loss caused by this discreteness. . In the spatial dimension, many features used in the detection, such as HOG, convolutional neural network features, etc., the stride when generating features is also greater than 1, and bounding box regression is also very helpful to the discreteness generated in this way. . In terms of detection speed, after using the outer border regression, more sparse sampling can be performed in the dimensions of space, scale, and aspect ratio, which can reduce the required computational cost.

2.2.4. Loss function

The convolution kernels used for target classification and positioning need to be trained. To design an end-to-end network model, the classification loss L_{cls} and positioning loss L_{reg} of the network need to be designed into a loss function. For the classification loss of the network, you can use the Softmax classification loss function.

Definition of the loss function of Faster R-CNN：

$$L(\{p_i\}, \{t_i\}) = \frac{1}{Ncls}\sum_i Lcls\,(p_i, p_i^*) + \lambda \frac{1}{Nreg}\sum_i p_i^* Lreg\,(t_i, t_i^*). \tag{2-5}$$

ISSN: 2664-9640

i is the index of the anchor in the small-batch,$p_i$is the probability that the anchor is the target. If the anchor is a positive sample, the probability of labeling the box p^* is 1, otherwise, it is 0.t_i is a vector representing the coordinate position, containing 4 parameters,t^* is the coordinate position vector of the label box.L_{cls} is a logarithmic classification function, only divided into target and background. For regression loss function $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$, R is a robust loss function smooth L1 (Girshick, 2015) $p_i^* L_{reg}$ Indicates that the loss function is calculated only when the anchor is a positive sample $(p^* = 1)$, Loss function when the anchor is a negative sample $p_i^* L_{reg}$ is $0(p^* = 0)$. These two loss functions are normalized by N_{cls} and N_{reg}respectively, where λ is the weight of balance. $L_{reg}$is calculated as:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & |x| \le 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{2-6}$$

This study found that there are a large number of packages in the express transit center in a stacked state. The mutual accumulation of packages makes the packages obstructed, which makes the target detection difficult to detect packages.

In the research of this article, it is found that Faster R-CNN has certain deficiencies in detecting stacked packages. It is easy to identify multiple packages as one package.

Figure 3 specific case

The purpose of package detection is to make the generated candidate frame closer to the real target frame to achieve the identification and detection of the package.

Package inspection often encounters dense packing of packages. For the problem of large piles of packages, this article uses RepGT loss (Wang et al., 2018)Adjusting the regression item in the package detection algorithm not only reduces the distance between the predicted frame and the corresponding real target frame but also increases the distance between it and the surrounding non-corresponding target frame (including the real target frame and the predicted frame). The loss function designed in this paper is as follows:

$$L = L(\{p_i\}, \{t_i\}) + L_{\text{RepGT}} \tag{2-7}$$

L_{RepGT} Represents the loss of the prediction box away from the surrounding target box with overlap. Assume $\rho_+ = \{P\}$ is the set of all candidate boxes representing positive samples (In this paper, the candidate frame with any target frame $IOU \ge 0.5$ is regarded as a positive sample candidate frame, and each target frame has at least one candidate frame), Assume $\varsigma = \{G\}$ is a collection of all target frames in an image. The purpose of L_{RepGT} loss is to keep the candidate frame away from the target frame of the non-specified target. Let B^p be the prediction frame obtained from the candidate frame p, G_{Attr}^P is the target frame with the maximum$IOU$of the candidate frame$p$, and G_{Rep}^P is the target frame with the maximum IOU of the candidate frame p (except for G_{Attr}^P). The calculation of $G_{Attr}^P$and G_{Rep}^P is as follows:

$$G_{Attr}^P = arg \max_{G \in \varsigma} IOU(G, P) \tag{2-8}$$

$$G_{Rep}^P = arg \max_{G \in \varsigma \setminus G_{Attr}^P} IOU(G, P) \tag{2-9}$$

B^p and G_{Rep}^P overlap calculation is defined as $IOG, IOG(B, G) \triangleq \frac{area(B \cap G)}{area(G)}$, The calculation of L_{RepGT} is as follows:

ISSN: 2664-9640

$$L_{\text{RepGT}} = \frac{\sum_{P\in\rho_+}^{\Sigma} Smooth\left(IOG\left(B^P, G^P_{Rep}\right)\right)_{ln}}{|\rho_+|} \tag{2-10}$$

$$Smooth_{ln} = \begin{cases} -ln(1-x) & x \le \sigma \\ \frac{x-\sigma}{1-\sigma} - ln(1-\sigma) & x > \sigma \end{cases}_{ln} \tag{2-11}$$

$Smooth_{ln}$ is a continuously differentiable function in the interval (0,1), parameter $\sigma \in [0,1)$ represents a parameter that is sensitive to outliers. It can be seen that if the candidate frame overlaps with a target frame that is not the current target, the RepGT loss function will increase the penalty loss, effectively preventing the predicted bounding box from moving to its non-target neighbors.

2.3. Experimental results

2.3.1. parameter settings

This article uses Faster R-CNN as a comparative experiment, adding the RepGT loss function L_{RepGT} to the loss function of Faster R-CNN. All experiments are carried out in the same experimental environment. The TensorFlow deep learning framework under the Ubuntu 16.04 system is used uniformly, a single NVIDIA RTX 2080Ti GPU is accelerated, and the stochastic gradient descent with a regularization coefficient (Weight Decay) of 10e-5 and momentum (Momentum) of 0.9 is used for optimization. The initial learning rate (Learning Rate) is 0.001, and after 30k iterations, the attenuation is 0.1 times, for a total of 100k iterations.

2.3.2. Result analysis

After adding the RepGT loss function, the surrounding target frame can be better separated, which is better than Faster R-CNN. It can be seen from the statistical average accuracy mAP that Faster R-CNN's package detection accuracy rate is 57.60, which reaches 59.98 after adding the RepGT loss function, which has a significant improvement inaccuracy.

Table I mAP

model	mAP
Faster R-CNN	57.60
Faster R-CNN +RepGT Loss	59.98

2.3.3. Ablation experiment

Statistics of experimental results for different parameters σ in the formula $Smooth_{ln} = \begin{cases} -ln(1-x) & x \le \sigma \\ \frac{x-\sigma}{1-\sigma} - ln(1-\sigma) & x > \sigma \end{cases}_{ln}$, The smoothing parameter σ is used to adjust the sensitivity of RepGT loss to non-target frames. When $\sigma = 1$, $Smooth_{ln}(1 - IOG)_{ln}$

Table II

Parameter σ	mAP	Promote
σ =0	59.52	1.92
σ =0.3	59.75	2.15
σ =0.5	59.40	1.80
σ =0.7	59.50	1.90
σ =1	59.98	2.38

It can be seen from the above table that the detection accuracy is the best when the parameter $\sigma = 1$ in the package detection data set is established in this paper, and there are still certain differences in the detection accuracy of the algorithm between different parameters.

ISSN: 2664-9640

3. Research on Package Detection Based on Feature Fusion

Aiming at the shortcomings of various basic feature extraction network frameworks, this chapter proposes a feature extraction network based on feature fusion and introduces the connection method of the network. A target detection algorithm is constructed using the new feature extraction network, and the specific execution process of the algorithm is analyzed in depth.

3.1. Multi-scale feature fusion

At present, multi-scale target detection is a research hotspot and focus in the field of target detection. In the early days of multi-scale detection research, most algorithms detected objects of different sizes in images by changing the form of sliding windows. At present, a better solution for multi-scale target detection is to use the image pyramid method (Liu et al., 2019). The image pyramid is a series of image sets with different resolutions ranging from fine to coarse according to certain rules. The bottom of the image pyramid is the original input image, so it has a higher resolution, but as the number of pyramid layers increases, The resolution of the image will gradually decrease.

This paper proposes a new feature extraction method based on multi-scale feature fusion. The main idea is to use low-resolution and high-semantic feature maps and high-resolution low-semantic feature maps to generate multiple Scale new feature maps, use the new feature map for subsequent target detection. The feature extraction network structure is shown in Figure 7. First, perform convolution operation on the input of the network, and then continue the convolution operation on the generated feature map. After repeated operations, feature maps of different resolutions are formed and then feature maps of different layers are used for fusion, and then Generate feature maps with relatively high semantics and high resolution, and finally use new feature maps with different resolutions and fusions for target detection.

3.2. Algorithm construction

The feature extraction method based on multi-scale feature fusion proposed in this section is a general method for constructing features when detecting multi-scale targets. This method is used to generate a feature map and a suggested area with the RPN network, and then a fully connected layer network is used for target detection. The original RPN network can be modified to build the feature extraction network based on multi-scale feature fusion proposed in this paper. It can also indirectly prove the simplicity and effectiveness of the feature extraction method.

The algorithm framework is mainly composed of the bottom-up network, top-down network, RPN network, and Fast R-CNN network. The bottom-up network structure generally refers to the basic feature extraction network. The top-down network structure is an extension of feature extraction. This network can be used to generate multi-scale feature maps. The proposed area is selected through the RPN network connected behind the network, and then the fully connected layer is used for target classification and detection.

4. Conclusion, limitations and further research directions

Deep learning has made remarkable achievements in computer vision. Target detection algorithms based on convolutional neural networks have made breakthrough progress in recent years. For package detection and recognition tasks, different detection frameworks reflect different performances. Therefore, according to the current common detection and recognition frameworks, this paper uses different feature extraction networks to verify the comprehensive performance of package detection and selects a network structure with better performance to lay a good foundation for the study of multi-scale package detection and

ISSN: 2664-9640

recognition systems. Although this article has done a lot of research and achieved certain results. However, there is still a lot of work to be improved in this research direction:

(1) Although the detection of small target packages has been improved to a certain extent, there are still some unsatisfactory, and the network scale used in the algorithm is large. In the actual application process, real-time performance is a big problem. How to improve the system Running speed without losing accuracy to a large extent is the next step of the research

(2) The scenario of the package data set is limited to the logistics transfer center. If there is more data to other scenarios, the trained model will be more robust.

References

[1] Cortes, C. & Vapnik, V. N. (1995), "Support-Vector Networks", *Machine Learning,* Vol. 20 No. 3, pp. 273-297.

[2] Erhan, D., Szegedy, C., Toshev, A. & Anguelov, D. (2014), "Scalable object detection using deep neural networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2147-2154.

[3] Felzenszwalb, Pedro, F., Girshick, Ross, B., McAllester, David, Ramanan & Deva (2010), "Object Detection with Discriminatively Trained Part-Based Models.", *IEEE Transactions on Pattern Analysis & Machine Intelligence,* Vol. 32 No. 9, pp. 1627-1645. Girshick, R. (2015), "Fast R-CNN", Computer Science,.

[4] Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014), "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", *in CVPR*, pp.

[5] Hoiem, Q. D. D. (2012), "Learning to localize detected objects", *in IEEE Conference on Computer Vision & Pattern Recognition*, pp.

[6] Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017), "Imagenet classification with deep convolutional neural networks", *Communications of the ACM*, Vol. 60 No. 6, pp. 84-90.

[7] Liu, Y., Wang, Y., Wang, S., Liang, T. & Ling, H. (2020), "CBNet: A Novel Composite Backbone Network Architecture for Object Detection".

[8] Sedaghat, A., Mokhtarzade, M. & Ebadi, H. (2011), "Uniform Robust Scale-Invariant Feature Matching for Optical Remote Sensing Images", *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 49 No. 11, pp. 4516-4527.

[9] Simonyan, K. & Zisserman, A. (2014), "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:*1409.1556,.

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015), "Going deeper with convolutions", *in Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9.

[11] Liu Yepeng,Wu Tongtong,Jia Xuejian&Zhai Yongjie.(2019), " A Multi-Scale Target Detection Method for Transmission Lines ", *INSTRUMENTATION, Vol. 026 No. 001*, pp. 15-18.