Application of Single Shot MultiBox Detector in Logistics Safety Testing

IOP Conf. Series: Materials Science and Engineering **569** (2019) 022002 doi:10.1088/1757-899X/569/2/022002

Application of Single Shot MultiBox Detector in Logistics Safety Testing

Xiaoling Xia[1], Xin Shi[1*], Qinyang Lu[1], Changqi Fan[2]

[1]College of Computer Science, Donghua University, Shanghai 201620, China

[2]i-Soft Co., Ltd, China

2181807@mail.dhu.edu.cn

Abstract. With the rapid development of the Internet economy, the load of the logistics industry has gradually increased, and people are paying more and more attention to the detection of restricted items in logistics packages. Using Convolution Neural Network can quickly and accurately identify the restrictions in the package, thus minimizing the pressure on personnel in logistics monitoring. This paper introduces the application method of SSD-based target detection model in the detection of logistics goods restriction products. Firstly, it analyses the characteristics of data sets and preprocesses the data and expounds the feasibility of SSD model in this respect. Different from the traditional upsampling method, this paper uses cubic spline interpolation to adjust the image resolution and expands the number of positive samples to balance the number of positive and negative samples. Next, batch normalization is added after the convolution layer and Softmax is replaced with multiple logistic classifiers so that the performance of the model can be improved. Finally, by comparing the results of SSD and Fast R-CNN, this paper proves that SSD, which obtains a mAp of 0.476 at more than 30 FPS, is better on this dataset. It owns both high accuracy and real-time speed.

1.Introduction

Target detection has always been an important direction in the field of computer vision research [1]. In many scenarios, it is necessary to extract the information in the image by means of target detection. As an important part of the daily parcel logistics industry and the security industry, the monitoring of restricted products covering X-ray films is responsible for preventing dangerous goods from entering the cargo channels, managing special cargo items such as knives, and monitoring the smuggling of national key contraband such as drugs. Traditional target detection methods such as HOG [2], SIFT [3], etc. are roughly divided into three parts, region selection, feature extraction, and classifier classification. This method requires artificial design features to achieve, which is not only difficult, but also the accuracy of the features cannot be guaranteed, so this method not only has low detection accuracy, but also is not robust. The invention of CNN makes the second and third parts of the traditional method can be combined.

Considering the real-time nature of the problem, this paper chose the SSD[4] model instead of the Faster R-CNN[5] model. While using the SSD model, some improvements were made to improve the performance of the model, and the results of the two models were compared at the end of the paper. In this paper, the dataset in the Tianchi Algorithm Challenge is used. This is a set of X-ray images of logistics packages. The resolution of the images is extremely uneven, as large as 1200*800, as small as 200*300. This makes it impossible to directly use the image of the dataset for the input of the SSD model. A better resize method is needed here, so that the image in the dataset can be resized to the same

resolution as clearly as possible. Compared with a variety of resize methods, this paper selects the cubic spline interpolation method to resize the image. In this paper, batch normalization is added and multiple independent logistic classifiers are used to replace the softmax classifier in the original model to improve the accuracy and speed of the model. In addition, data enhancement can significantly improve the performance of the algorithm. For the package image of the unlimited product in the training set, it cannot be directly used for the input of the model. However, in the case where the data set itself is small, it is proposed in this paper to extract the limited product and expand the dataset by aiming at the target in the image and overwriting the normal image to increase the amount of data. Expanding the data set can improve the accuracy of the model to some extent.

2.Related work

2.1. SSD

The VGG16 is used as the basis of the network structure, and 8 convolutional layers are added to the VGG16 structure to obtain multi-scale feature maps for target detection. Low levels are used to predict large size targets, and high levels are used to predict small size targets. SSD combines the regression idea of YOLO [6] model with the anchor mechanism of Faster R-CNN. This not only makes the calculation of the neural network simple, but also makes the detection more real-time. The loss function of the SSD is divided into two parts, and the confidence loss of the corresponding default box and the target category and the corresponding position regression are calculated.

2.2. Faster R-CNN

Faster R-CNN inputs the entire picture into the CNN network for feature extraction, generates the proposal with RPN, maps the proposal to the last feature map of the convolutional network, and generates a fixed-size feature map through each Rol through the Rol pooling layer. Faster R-CNN utilizes Softmax Loss and Smooth L1 Loss for joint training of classification probabilities and Bounding tree regression.

2.3. YOLO

After inputting the image, YOLO is responsible for extracting features through the convolution part. The full connection part is responsible for prediction, and then filtering the bounding box by non-maximum value suppression. Finally, the remaining bounding box is the result of the target detection. YOLO first divides the original image into a grid. If the center point of an object falls on a cell, the cell is responsible for predicting the object. Each cell needs to predict the value of multiple borders, and predicts a set of confidence scores for each border. YOLO's loss function is divided into three parts, the loss of the border, the loss of confidence and the loss of classification. YOLO uses GoogleNet, which is faster than VGG-16. YOLO only uses an 8.52 billion operation for a forward process, while VGG-16 requires 30.69 billion, but YOLO’s accuracy is slightly lower than VGG-16.

3.Method

3.1. Model

In 2014, Girshick et al. proposed the use of region proposal + CNN, followed by R-CNN [7], Fast R-CNN [8], Faster R-CNN, mask R-CNN [9] and other models. In 2015, Redmon J proposed the YOLO model. YOLO can achieve a frame rate of 45fps in speed. In 2016, WLiu et al. proposed the SSD model, which achieved good results in both detection accuracy and detection time. These models can be divided into one-stage and two-stage modes. The former means that the bounding box of the extracted target is no longer separated from the classified target, and the latter has a step of region proposal, that is, using the RPN network. The regression processing is performed on the bounding box, and the bounding box is found before the object is classified and predicted. The one-stage model is characterized by fast speed, while the two-stage model dominates the accuracy and accuracy of the detection.

Faster R-CNN first uses RPN to generate proposals. The complexity of the calculation greatly affects the speed of the whole prediction process. Even if the accuracy is high, it does not perform well in real-time. In this paper, there are lighters, scissors, knives, etc. among the restrictions that need to be detected. The size of the lighter is small, and the size of the tool is large, which means that the size of the target is different. The SSD model consists of VGG16 and multiple convolutional layers. The convolutional layer can output multiple feature maps to detect targets of different scales. YOLO method model training relies on object recognition annotation data, so YOLO detection is not ideal for unconventional object shapes or proportions. The SSD adopts the multi-scale feature map method. Region candidate frames of different sizes and different aspect ratios are set on different scale feature maps. The area candidate frame is defined as follows. The SSD will set regional selection boxes of different sizes and aspect ratios on different scale feature maps. The area candidate box is defined as formula (1). Where m is the number of feature layers, S_{min} is the lowest feature layer scale, S_{max} is the highest feature layer scale, and the intermediate feature layer scale is evenly distributed.

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m-1}(k-1) \ , k \in [1, m] \tag{1}$$

In this paper, the SSD model is selected for prediction, and some improved methods are added to expect better prediction results. And faster calculation speed. The structure of the SSD model is given in Figure 1.

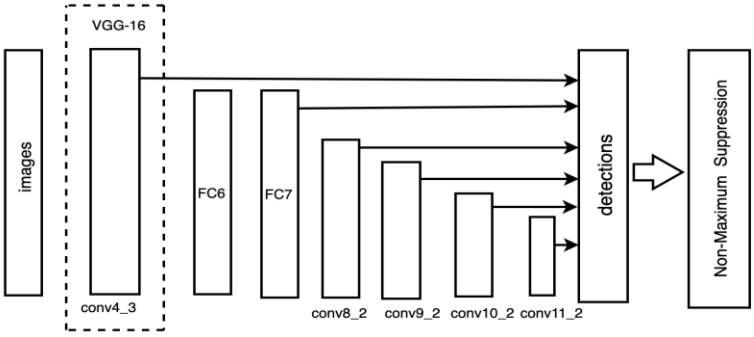

Figure1.SSD model network architecture

3.2. Optimization and improvement

3.2.1. Resize

The SSD model needs to input the same length and width. The first thing we need to do is the resize of the image size. We need a suitable method to avoid the loss of the image features and the resolution of the image. In the dataset used in this paper, the resolution of the image varies greatly. To unify the resolution of different images, it is necessary to enlarge the image of the small resolution and reduce the image of the large resolution. In the traditional method, the upsampling method is used to deal with this problem. In this paper, we try to use the cubic spline interpolation method to deal with the resolution problem of the image. This method can guarantee the picture features in the resize process more than the traditional upsampling method. The cubic spline interpolation uses the 4×4 neighbourhood pixel double 3rd interpolation, and the interpolation result of all points is shown in the formula (2), where *B (x, y)* is the target pixel and W is the weight.

$$B(X,Y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij} \times W(i) \times W(j) \tag{2}$$

3.2.2 Sample expansion

In the various methods of deep learning, the sample is often expanded to obtain better training results. Faster R-CNN's approach to train sample expansion is to flip the image horizontally with a 50% probability. In addition to using horizontal flipping, the author of SSD randomly crops the images in the training set according to the probability of 50%, and proposes this operation in the paper to improve the effect of the training model. In order to maximize the size of the data set and improve the accuracy of the prediction results, the data set has been augmented in this paper. The imbalance pictures between positive and negative samples is a huge waste of the originally small data set, so a method is needed to make full use of the target. To some extent, the expanded training sample library can improve the accuracy of detection. In this paper, the object in the 981 pictures of the limited product is cut out according to the position of the target boundary frame in the data set and is attached to the picture without the target according to the original position, in this case, it increase the number of pictures used in this data set and provide us more testing data set.

3.2.3. Classifier

YOLOv3[10] replaces softmax with multiple independent logistic classifiers, and the accuracy does not decrease. In this paper, we try to apply this method to the SSD model, and use the logistic classifier instead of softmax to improve the speed of the model. formula (3) is the softmax classifier and formula (4) is the logistic classifier.

$$h_\theta\left(x^{(i)}\right) = \begin{bmatrix} p\left(y^{(i)} = 1 \middle| x^{(i)}, \theta\right) \\ p\left(y^{(i)} = 2 \middle| x^{(i)}, \theta\right) \\ \vdots \\ p\left(y^{(i)} = k \middle| x^{(i)}, \theta\right) \end{bmatrix} = \frac{1}{\sum_{c=1}^{k} e^{\theta_c^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix} \tag{3}$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

3.2.4. Batch normalization.

In YOLOv2[11], a normalized pre-processing is done for each batch of data. By adding batch normalization after each convolutional layer, this method greatly improves the slow convergence of the model, and reduces the dependence on other regular methods with increased the accuracy of the model. In this paper, we also tried to apply batch normalization method. The gradient disappearance problem of deep neural networks often leads to slower convergence when training deep neural networks, and batch normalization is to force the more and more biased distribution back to the more standard distribution through certain normalization methods which active the input. The value falls in the area where the nonlinear function is sensitive to the input, so that small changes in the input result in a large change in the loss function. This not only improves the training speed, but also greatly speeds up the convergence process and increases the classification effect.

4. Experiments

4.1. Experimental result

Figure 2 shows the test results of the partially wrapped samples using the SSD model, and Figure 3 shows the results of the corresponding faster R-CNN model. In Figure 2, the bounding box of the tool is more precise than the bounding box of the tool in Figure 3. The scissors identified in Figure 2 were not detected in Figure 3. As can be seen from the picture, in some cases, the accuracy of the SSD model is even better than the faster R-CNN.

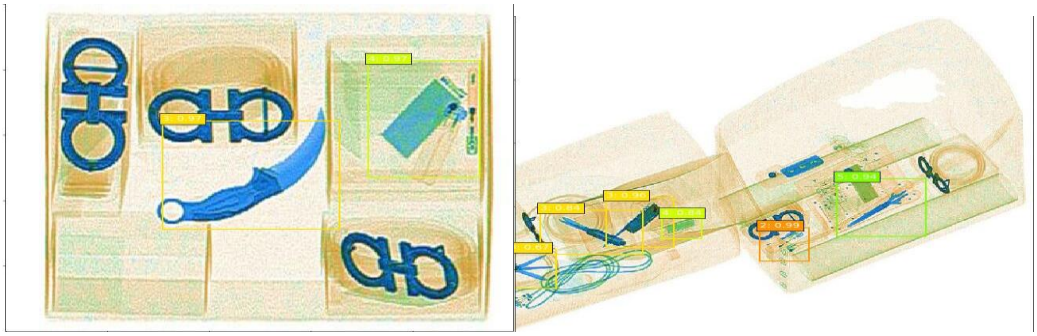

Figure 2. Example of SSD model test results

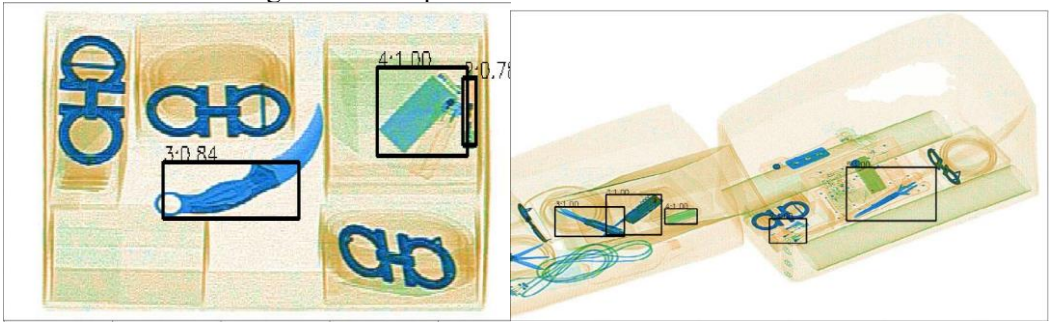

Figure 3. Example of Faster R-CNN test results

4.2. Parameter

The AP is the area under the Precision-Recall curve. For the two-class problem, the sample can be divided into four cases: true positive, false positive, true negative, and false negative according to the combination of its real category and the learner prediction category, so that TP, FP, TN, and FN respectively represent their corresponding number of samples, and TP + FP + TN + FN represents the number of samples. Precision and Recall are defined as formula (5) and (6), respectively.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

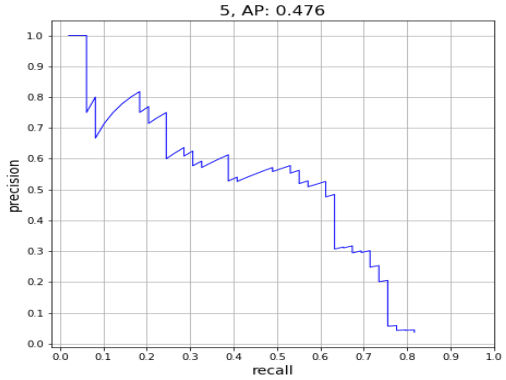

Figure 4. Precision-Recall curve for the scissors category

Figure 4 shows the Precision-Recall curve of the scissors category. Taking the average value of APs of each category, the comprehensive index mAP, Mean Average Precision, is obtained, and the mAP index is used to evaluate the performance accuracy of the target detection model. The experimental mAP is not as good as Faster R-CNN. However, from the training parameters, the more compact the model

has, the lower the training parameters are, and the training parameters of SSD are almost one-third of Faster R-CNN.

5. Conclusion

In this paper, we used the SSD model to perform limit detection on wrapped X-rays and found some ways to increase the performance of the model. Resetting the resolution of the image using cubic spline interpolation is useful for saving the features of the image. Extracting the target in the image sample with the target and cutting it into the sample without the target can make full use of the limited data set and balance the positive and negative samples. By adding Batch Normalization to the model, the input of the neural network of each layer maintains the same distribution. Although the loss at the beginning shows a large value, the value of loss is significantly faster than the previous rate of decline, and finally converges. Using multiple logistic classifiers instead of softmax reduces the amount of computation and does not degrade the accuracy of the model. From the results, although the overall detection accuracy of the model is not comparable to Faster R-CNN, the accuracy is also improved on the basis of guaranteed speed.

References.

[1] Erhan D, Szegedy C, Toshev A, et al. (2014) Scalable object detection using deep neural networks. In: Proc. Of the IEEE Conf. On Computer Vision and Pattern Recognition. New York. pp. 2147-2154.
[2] He N, Cao J, Song L. (2008) Scale space histogram of oriented gradients for human detection. In: Information Science and Engineering, 2008. ISISE'08. International Symposium on. IEEE. New York. pp. 167-170.
[3] BAY H, ESS A, TUYTELAARS T, et al. (2008) Speeded-up robust features (SURF). Computer vision and image understanding., 110 (3) :346-359.
[4] Liu W, Anguelov D, Erhan D, et al. (2016) SSD: single shot multibox detector. In: European Conference on Computer Vision. Springer International Publishing. New York. pp. 21-37.
[5] Ren S, He K, Girshick R, et al. (2015) Faster r-cnn: Towards real time object detection with region proposal networks. In: Advances in Neural Information Proc. Systems. New York. pp. 91-99.
[6] Redmon J, Divvala S, Girshick R, et al. (2016) You only look once: Unified, real-time object detection. In: Proc. Of the IEEE Conf. On Computer Vision and Pattern Recognition. New York. pp. 779-788.
[7] Girshick R, Donahue J, Darrell T, et al. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. Of the IEEE Conf. On Computer Vision and Pattern Recognition. New York. pp. 580-587.
[8] Girshick R. (2015) Fast R-CNN. In: Proc. Of the IEEE Inter. Conf. On Computer Vision. New York. pp. 1440-1448.
[9] Kaiming H, Georgia G, Piotr D, et al. (2017) Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV). New York. pp. 2980-2988.
[10] Redmon J, Farhadi A. (2018) YOLOv3: An Incremental Improvement. In: IEEE Conference on Computer Vision and Pattern Recognition. New York. pp. 89-95.
[11] Redmon J, Farhadi A. (2017) YOLO9000: Better, Faster, Stronger. In: IEEE Conference on Computer Vision and Pattern Recognition. New York. pp. 6517-6525.