Deconvolution Single Shot MultiBox Detector for Supermarket commodity detection and classification

Dejian Li *[a], Jian Li[b], Binling Nie[a] , Shouqian Sun[a]
[a]College of Computer Science and Technology, Zhejiang University, Hangzhou, China; [b]Nanjing University of Science and Technology, Nanjing, China
dejianli@zju.edu.cn, lijiannuist@gmail.com, nbl1221@zju.edu.cn, ssq@zju.edu.cn

ABSTRACT

This paper proposes an image detection model to detect and classify supermarkets shelves' commodity. Based on the principle of the features directly affects the accuracy of the final classification, feature maps are performed to combine high level features with bottom level features. Then set some fixed anchors on those feature maps, finally the label and the position of commodity is generated by doing a box regression and classification. In this work, we proposed a model named Deconvolutiuon Single Shot MultiBox Detector, we evaluated the model using 300 images photographed from real supermarket shelves. Followed the same protocol in other recent methods, the results showed that our model outperformed other baseline methods.

Keywords: image detection, image classification, deep learning

1. INTRODUCTION

In the last decades, shopping via supermarket has greatly integrated into people's lifestyle. And with thousands of commodities crowded into supermarket shelves, how to record commodity's status in time is became a crucial issue. Based on commodity inquiry system to manually record commodity's status is a feasible way. However, this costs lots of manual labor and is less efficient. Furthermore, on the one hand, with extensive use of electronic computers and other modern equipment, it is very convenient to collect lots of data, however, only through storing these data in the dataset and make statistical analysis, these data can become valuable.

At the same time, image plays an important role in human acquisition and exchange information. And more and more image related technology is applied in the stuff of social life. Based on this trend, Cognitive Context Reasoning for Computer Vision (CCRVC2016[1]) held a competition to discuss image related computer vision topics including intelligent supermarket management, and the goal of intelligent supermarket management starts from detect and classify supermarket commodity.

Detection has been a well-researched problem in the computer vision community in the last decades. Following the common pipeline of "region selection + feature selection + classifiers" in object detection. One of the earliest structure to achieve not bad result was adopted sliding window with different length to width ratio to iterate whole image to get region of interests (ROIs), then feed those ROIs in a manually design image feature selector (e.g. SIFT [2], Hog [3]), and then feed into classifiers to get results. Particularly, the extraction of the features directly affects the accuracy of the final classification. A variety of recent papers provide methods for generating category independent region proposes. Such as objectness [4], selective search [5], multi-scale combinatorial grouping [6] and so on. Since Ross et.al [7] made impressive results on PASCAL VOC, COCO, and ILSVRC detection tasks through applied high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects. After that, this type of structure has prevailed on detection tasks, for example, Kaiming et. al [8] added a spatial pyramid pooling layer on the top of last convolution layer to generate fixed length outputs. Furthermore, Shaoqing et.al [10] redesigned region selection part called RPN (region proposal networks) to directly generate region proposal by sliding last convolution layer once.

On the other hand, consider the real-time application needs and computability of embedded system, Joseph et.al [11] provided a fancy idea which accelerate detection speed by treating the whole task as a regression problem. And later Wei

et.al [12] combined Joseph's regression idea with Shaoqing's anchor mechanism shows the real possibility of target detection in practical application.

Different from the existing detection models, this paper proposes a more elegant architecture, the so-called Deconvolution Single Shot Multibox Detector (Deconv. SSD). Then fusion them into a supermarket detection model as illustrated in Fig 1. Our main contribution lies in the following two aspects. First, this work is a preliminary study of intelligent supermarket management which not only has a high research values but also has a practical import. Second, we modify and extend basic single shot multibox detector such that it works with few training images and yields more precise detections. The main idea is to supplement a usual network by deconvolution layers, these increase the resolution of the output. The results showed that the new detection classifier accomplished a satisfied speed and precision.

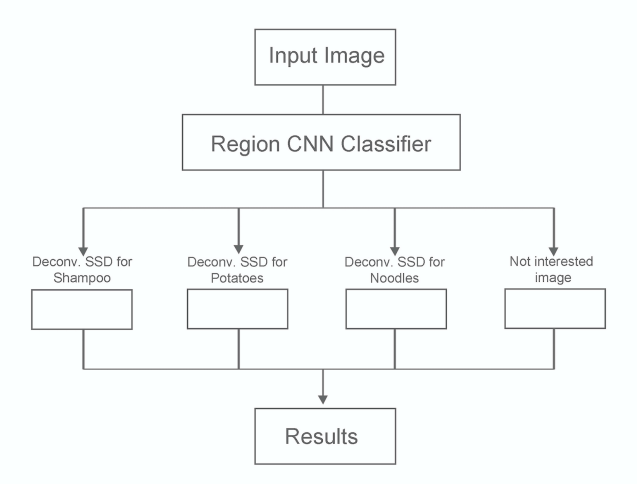

Figure 1. Our architecture of the fusion Supermarket detection model.

The rest of this paper is organized as follows. Section 2 details the proposed supermarket detection framework and Deconv. SSD model. Experimental results and analysis are presented in Section 3. Finally, conclusions and future works are given in Section 4.

2. THE SUPERMARKET DETETECT MODEL

In this work, we redesigned a SSD model to minimize information loss when rescaling image into a small size to decrease preprocessing time. Our fusion model made assumption that each area only occupies its category commodity, this is make sense because it is in line with the user's shopping habits.

2.1 Region classify model

We extract a 4096-dimensional feature vector from each input image using Caffe [13] implementation of the CNN described by Krizhevsky et al. [9]. Features are computed by forward propagating through five convolutional layers and two fully connected layers. We detailed the architecture sizes in table 1. Layers 1-5 are almost the same as AlexNet architecture but with the following difference. We removed contrast normalization part and used a smaller stride (2 instead 4) for a larger feature maps. For the super-variables, we used asynchronous stochastic gradient descent with 0.9 momentum, fixed learning rate schedule (learning rate=1e-2, decreasing the learning rate by 10% every 100 epochs). The results show we can achieve 95.4% accuracy, and as for wrong perdition, the reason is because the original images are blur, we guess this phenomenon is caused by camera focus is not completed on the shot.

2.2 Deconv. SSD model

While YOLO was the first used regression idea to directly got object proposal from input image, and through this fancy idea, the detection speed is good enough for real time application. But its accuracy still needs to be improved. The main

weak part in YOLO is it only predict two boxes for each label in each area. To improve this, SSD use of multi-scale convolutional bounding box outputs attached to multiple feature maps at the top of the network. This adjustment allows SSD efficiently model the space of possible box shapes, and its performance allows us to see the real possibility of target detection in practical applications. Therefore, we redesigned basic SSD model to a new model called Deconv. SSD model. Figure 2 shows details and modification of our Deconv. SSD model.

Table 1: Architecture specifics for region classify model. The spatial size of the feature maps depends on the input image size. Here we show one scale of our training spatial sizes.

Layer	1	2	3	4	5	6	7	Output
Stage	conv+max	conv+max	conv	conv	conv+max	full	full	full
#channels	96	256	512	384	256	4096	4096	4
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	2x2	1x1	1x1	1x1	1x1	-	-	-
Pool size	3x3	2x2	-	-	2x2	-	-	-
Pool stride	2x2	2x2	-	-	2x2	-	-	-
Padding size	-	-	1x1x1x1	1x1x1x1	-	-	-	-
Spatial size	501x501	82x82	39x39	39x39	39x39	7x7	1x1	1x1

In order to deal with high resolution image and multiple commodity types, one simple way is to enlarge image size so that detector can better localize object target. However, this will sharply increase the occupation of calculation resources and slow the process speed, which is not meet the requirements of real time application. Therefore, our group came up an idea to resample network middle generate features along with reduce image resolution, this could offer a lot of benefits such as decrease GPU memory space and improve the speed of training. Furthermore, we also made some adjustment about the relevant parameters of the algorithm. Since each commodity type hardly appears to be coincided, we set none maximum suppression (nms_threshold) to 0.2. As for count multibox loss proposals, we set both overlap threshold and negative overlap to 0.2, and increases an aspect ratio 4.

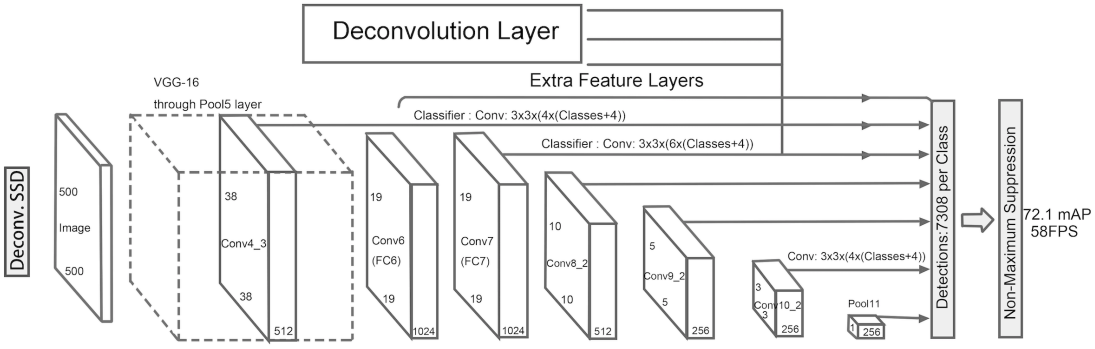

Figure 2. Structure of Deconvolution SSD framework.

2.3 Measurement

The detection task will be judged by the precision/recall curve. The principle quantitative measure used will be the average precision (AP) and mean average precision (MAP). Detections are considered true or false positives based on the area of overlap with ground truth bounding box. To be considered a correct detection, the area of overlap a_0 between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 50% intersection over union (IOU).

Particular, for CCRVC2016 dataset, if the shelf areas are not contained full commodity in horizontal or vertical then this area do not need to be counted into detection, otherwise, the detection area need to be extended align into shelf boundary (See Figure 3).

When consider the real time detection application, often detection speed is measured in frames per second (FPS).

Figure 3. Samples of effective areas. The red cross in the top left picture shows example of area which do not need to be measured. On the contrary, the green circle in the top right shows effective area.

3. DATA AUGUMENT AND RESULTS

3.1 The CCRVC2016 dataset

In this work, we joint CCRVC2016 computer vision competition for supermarket commodity detection task, and our model achieved top one in the competition. The supermarket dataset contains 1000 jpg files, those files contained more than 300 commodity brand in three categories of products (instant noodles, potato chips, shampoo). All pictures are taken from real supermarket and contains zero or more target commodity. Follow the measurement in Section 2, whole dataset offers a total of 6211 bounding boxes. However, few of them were 'noisy' boxes – for instance, some images dpi error cause coordinate conversion exception exceeds image boundary, which make those boxes are proper boxes. To filter them, we wrote a small program and automatically removed them from the dataset. This process left us a total of 5319 bounding boxes. Then finally we got instant noodle boxes (2265, 42.58%), potato chips boxes (1656, 31.13%) and shampoo boxes (1398, 26.29%).

3.2 Data augmentation

Data augmentation is essential to teach the network the desired invariance and robustness properties, when only few training samples are available. Worth to mention, we participated in the CCRVC2016 supermarket competition with no external data used for training. As for total 240 commodities types, we adopted a set of techniques for data augmentation, which we describe next.

For an original training image, we cropped 1/3 of left, center, right and bottom and take the 2/3 of rest image, which gives us total five images. For the rest four 1/3 images, we did stitching operation, for each 1/3 image, we stack it above the rest images. Worth to mention, bounding boxes should be recalculating for those augment images. And those data augment is within our assumption, that is to say that we only did crop through same region images.

3.3 Results

We evaluated our model by comparing with same setting in SSD and default setting in YOLO. Table 2 shows the results. Note all models are trained with same data augment method descripted above.

The results show that, the use of default SSD framework produced higher accuracy and the fastest speed among three (68.8 vs. 62.5, 59 vs. 58), which is also consistent with the improvements made by SSD. Because of generate bounding box proposals and feature resampling stage, YOLO has the slowest speed (44 vs. 58). When eliminating these and using multiple layers for prediction at different scales instead, our Deconv. SSD lost a little bit of speed in exchange for a substantial increase in accuracy (72.1 vs. 68.8).

Table 2. Speed and Accuracy by different detection frameworks

Detection Framework	MAP	FPS
SSD[12]	68.8	59
YOLO[11]	62.5	44
Decov. SSD	72.1	58

4. CONCLUSION AND FUTURE WORK

In this paper, we presented our detection model for detect commodities in supermarket shelves. Our model is an ensemble of Decov. SSD and achieve the top in the CCRVC2016 competition. Our model is also suitable for real time application (achieved 58FPS), 0.72 map scores showing the possibility to deploy our model into real supermarket. For future work, we plan test and deploy our model in the supermarket, learning to investigate the possible performance gain by doing so.

REFERENCES

[1] Cognitive Context Reasoning for Computer Vision (CCRCV 2016)

[2] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 886-893.

[4] Alexe B, Deselaers T, Ferrari V. Measuring the objectness of image windows[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2189-2202.

[5] Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[6] Arbeláez P, Pont-Tuset J, Barron J T, et al. Multiscale combinatorial grouping[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 328-335.

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[8] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision. Springer International Publishing, 2014: 346-361.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[J]. Conference on Computer Vision and Pattern Recognition(CVPR), 2016.

[12] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. arXiv preprint arXiv:1512.02325, 2015.

[13] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

AUTHORS' BACKGROUND

Your Name	Title*	Research Field	Personal website
Dejian Li	Phd candidate	Image Detection, deep learning	
Jian Li	Master student	Image Detection, deep learning	
Binling Nie	Phd candidate	Data mining, deep learning	
Shouqian Sun	full professor	Multimedia and vision, Human-computer interaction	

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor