Expert Systems With Applications 113 (2018) 1–17

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A Multi-strategy Region Proposal Network

Yu-Peng Chen[a,b], Ying Li[a,b], Gang Wang[a,b,*], Qian Xu[a,b]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, People's Republic of China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, People's Republic of China

ARTICLE INFO

Article history:
Received 4 January 2018
Revised 18 June 2018
Accepted 19 June 2018
Available online 28 June 2018

Keywords:
Object detection
Region proposal generation
Convolutional neural network
Classification

ABSTRACT

The Faster Region-based Convolutional Network (Faster R-CNN) was recently proposed achieving outstanding performance for object detection. Specially, a Region Proposal Network (RPN) is designed to efficiently predict region proposals with a wide range of scales and aspect ratios in Faster R-CNN. Nevertheless, once the number and quality of region proposals generated by RPN are not ideal the object detection performance of Faster R-CNN is affected. In this paper, multiple strategies are applied to address these limitations and improve RPN. Hence, a novel architecture for region proposal generation is presented which is named as Multi-strategy Region Proposal Network (MSRPN). Four improvements are presented in MSRPN. Firstly, a novel skip-layer connection network is designed for combining multi-level features and boosting the ability of pooling layers. Thereupon, the quality of region proposals is strengthened. Secondly, improved anchor boxes are introduced with adaptive aspect ratio and evenly distributed interval of selected scales. In this way, the number of predicted region proposals for detection is seriously reduced and the efficiency of object localization is increased. Particularly, the capability of small object detection is enhanced by applying the first and second improvements. Thirdly, classification layer and regression layer are unified as a single convolutional layer. Furthermore, the model complexity of output layer is reduced. Thus, the speed of training and testing is accelerated. Fourthly, the bounding box regression part of multi-task loss function in RPN is improved. Consequently, the performance of bounding box regression is promoted.

In the experiment, MSRPN is compared with the Fast Region-based Convolutional Network (Fast R-CNN), Faster R-CNN, Inside-Outside Net (ION), Multi-region CNN (MR-CNN) and HyperNet approaches. MSRPN achieves the state-of-the-art mean average precision (mAP) of 78.9%, 74.8% and 32.1% on PASCAL VOC 2007, 2012 and MS COCO data sets with the deep VGG-16 model, surpassing other five object detection methods. Simultaneously, the above experiment results are obtained by MSRPN with only 150 region proposals per image. Additionally, MSRPN gets excellent performance on small object detection. Furthermore, MSRPN runs at 6 fps which is faster than other methods. In conclusion, the MSRPN method can provide important support for the intelligent object detection systems.

1. Introduction

Over the years, the problem of object detection (Alexe, Deselaers, & Ferrari, 2010; Erhan, Szegedy, Toshev, & Anguelov, 2014; Ren, & Ramanan, 2013; Yuting, Kihyuk, Ruben, Gang, & Lee, 2015) for images has been studied deeply but still remains a challenging task. In computer vision, the methods for solving object detection problem can be divided into two categories: the methods based on hand-engineered feature (Felzenszwalb, & Huttenlocher, 2000; Schmid, & Mohr, 1997; Weber, Welling, & Perona, 2000) and the methods based on deep learning network (Carreira, & Sminchisescu, 2010; Cheng, Zhang, Lin, & Torr, 2014; Ghodrati, Pedersoli, Tuytelaars, Diba, & Gool, 2015; Kuo et al., 2015; Zitnick et al., 2014). The well known methods based on hand-engineered feature include Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Histograms of Oriented Gradient (HOG) (Dalal, & Triggs, 2005) and Deformable Part Models (DPM) (Felzenszwalb, Girshick, McAllester, & Ramanan, 2010). SIFT is a method for extracting distinctive invariant features from images. These features can be used to perform reliable matching between different views of an object or scene. HOG shows that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. DPM applies im-

[*] Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun 130012, People's Republic of China.
E-mail addresses: wanggang.jlu@gmail.com, quarryrock@163.com (G. Wang).

https://doi.org/10.1016/j.eswa.2018.06.043
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

age descriptors such as HOG or SIFT features and sweeps through the entire image to find regions with a class-specific maximum response.

Recently, methods based on deep learning network have achieved a great success in object detection. By comparison with the traditional methods, object detection methods like Over-Feat (Sermanet et al., 2014) and Multi-column deep neural networks (Ciresan et al., 2012) have showed dramatic improvements in accuracy and speed. Particularly, in traditional methods such as DPM and HOG, the generalization of object recognition models is poor, once they strongly depends on the prior knowledge of the designer respect to the target object and image features. Nevertheless, methods based on deep learning network such as ConvNets (Krizhevsky, Sutskever, & Hinton, 2012) may perform better regarding generalization because ConvNets is suitable for various objects detection with only one model. Specially, the localization and classification accuracy of methods based on deep learning network are strengthened by applying high quality region proposals (Hosang, Benenson, Dollár, & Schiele, 2015; Hua et al., 2015). The current research status of primary methods for object detection based on deep learning network is described as follows.

In recent years, excellent detection results are achieved by Region-based Convolutional Neural Network (R-CNN) (Girshick, Donahue, Darrell, & Malik, 2015). Moreover, the Fast R-CNN (Girshick, 2015) is designed to promote the performance of object detection based on R-CNN. Furthermore, the computation bottleneck of region proposals in Fast R-CNN is solved by Faster R-CNN (Ren, He, Girshick, & Sun, 2017) which contains two stage of operations. First, RPN is applied to replace the Selective Search (SS) (Uijlings, van de Sande, Gevers, & Smeulders, 2013) method for predicting region proposals. Second, the predicted region proposals are used by Fast R-CNN to detect object. The RPN is a kind of fully convolutional network (FCN) (Long, Shelhamer, & Darrell, 2015) and can be trained end-to-end for generating detection region proposals.

In the works mentioned above, four problems can be identified. Firstly, lower and higher features are not efficiently applied to generate region proposals. Especially, only max pooling method is used to extract features. Therefore the quality of region proposals is not fine. Secondly, the aspect ratio of each anchor box in RPN is fixed. In addition, the interval of selected scales is not evenly distributed. Consequently, the number of predicted region proposals for detection is too much and the ability of localization is relatively poor. Thirdly, the classification layer and regression layer are separated in RPN. Specially, the model complexity of output layer is large. Thereupon, the speed of training and testing is degraded. Fourthly, the regression offsets are not reasonably expressed in bounding box regression formulation of RPN. As a result, the bounding box regression performance of RPN is affected.

In this paper, a novel Multi-strategy Region Proposal Network (MSRPN) for region proposal generation is proposed. Four corresponding improvements are presented in MSRPN. Firstly, a novel skip-layer connection network is designed for combining multi-level features and boosting the ability of pooling layers. Thereupon, the quality of region proposals is strengthened. Secondly, improved anchor boxes are introduced with adaptive aspect ratio and evenly distributed interval of selected scales. In this way, the number of predicted region proposals for detection is seriously reduced and the efficiency of object localization is increased. Particularly, the capability of small object detection is enhanced by using the first and second improvements. Thirdly, classification layer and regression layer are unified as a single convolutional layer. Furthermore, the model complexity of output layer is reduced. Thus, the speed of training and testing is accelerated. Fourthly, the bounding box regression part of multi-task loss function in RPN is improved. Consequently, the performance of bounding box regression is promoted.

Section 2 presents the related work to our MSRPN. Section 3 shows the basic concept of Faster R-CNN. Section 4 presents the improved MSRPN method and its implementation details. Section 5 describes the experiment results and discussions. Finally, section 6 draws some conclusions for this paper.

2. Related work

In this section, we review the current deep leaning based methods most related to our work. The mean average precision (mAP) has been improved based on R-CNN (Girshick et al., 2015).The gap between image classification and object detection is bridged by R-CNN. Three processes are contained in R-CNN: First, around 2000 category-independent region proposals are generated by applying SS method. Second, a fixed-length feature vector from each region is extracted by a large pre-trained CNN (Krizhevsky et al., 2012; LeCun et al., 1989; Torralba, 2003). Third, a set of linear Support Vector Machines (SVMs) (Chen et al., 2017) is used to classify the feature vector. Nevertheless, the complexity of R-CNN is increased by applying this multi-stage pipeline.

Measures are taken by Fast R-CNN (Girshick, 2015) to promote the performance of object detection based on R-CNN. Fast R-CNN is a single-stage training method that jointly learns to classify region proposals and refine their spatial locations. Moreover, the region of interest (RoI) pooling strategy based on the top level feature is more efficient than the R-CNN feature extracting method. Multi-task training of Fast R-CNN is convenient because it avoids managing a pipeline of sequentially-trained tasks. Fast R-CNN applies SS method to generate region proposals. However, the detection speed is reduced by the region proposal generation step. The computation bottleneck of Fast R-CNN is solved in Faster R-CNN (Ren et al., 2017). Nevertheless, the top level features of Faster R-CNN are too coarse for object detection and classification. Therefore, the ability of object detection still needs to be improved.

MR-CNN (Gidaris, & Komodakis, 2015) introduces a bounding box regression scheme to improve results on VOC, where bounding boxes are evaluated twice. This method depends on a multi-region deep CNN that encodes semantic segmentation-aware features. Nevertheless, the region proposals of MR-CNN are generated by SS. In addition, it is time-consuming to evaluate additional boxes and to add semantic segmentation results. ION (Bell, Zitnick, Bala, & Girshick, 2016) is an object detector that exploits information both inside and outside the region of interest. Skip pooling (He, Zhang, Ren, & Sun, 2015) is used to extract information at multiple scales and levels of abstraction. However, the region proposals generation method for ION is not efficient. HyperNet (Kong, Yao, Chen, & Sun, 2016) is designed for handling region proposal generation and object detection jointly. Good object detection accuracy is obtained by HyperNet on PASCAL VOC 2007 and 2012. Nevertheless the loss function for HyperNet needs to be improved. Among these methods, the proposed MSRPN is compared with the Fast R-CNN, Faster R-CNN, ION, MR-CNN and HyperNet approaches. From the experiment results, we can see that the performance of our MSRPN is better than other five object detection methods.

3. Basic concept of Faster R-CNN

Faster R-CNN method is designed as a unified system for object detection. Section 3.1 introduces the architecture of RPN. Section 3.2 presents the training steps of Faster R-CNN. The parameters of Faster R-CNN are showed in Table 1.