Deep Convolutional Neural Network for Pedestrian Detection with Multi-Levels Features Fusion

Danhua Li[1], Xiaofeng Di[1,a], Xuan Qu[1], Yunfei Zhao[1] and Honggang Kong[1]

[1]China Academy of Transportation Sciences, 100088 Beijing, China

Abstract. Pedestrian detection aims to localize and recognize every pedestrian instance in an image with a bounding box. The current state-of-the-art method is Faster RCNN, which is such a network that uses a region proposal network (RPN) to generate high quality region proposals, while Fast RCNN is used to classifiers extract features into corresponding categories. The contribution of this paper is integrated low-level features and high-level features into a Faster RCNN-based pedestrian detection framework, which efficiently increase the capacity of the feature. Through our experiments, we comprehensively evaluate our framework, on the Caltech pedestrian detection benchmark and our methods achieve state-of-the-art accuracy and present a competitive result on Caltech dataset.

1 Introduce

Pedestrian detection has been exhaustively explored in the last decade because of its growing importance in real-world applications, such as automatic driving, human behavior analysis [1] or intelligent video surveillance [2]. Despite the great progress has been made by deep convolutional neural network on pedestrian detection, recent papers still show significant improvements, suggesting that a saturation point has not yet been reached [3].

Pedestrian detection methods can be simply divided into two categories: The first is to use traditional pedestrian detection method which needs manual design to extract features for each proposal and classify them by a trainable classifiers, common methods for pedestrian detection task is to use sliding window based techniques for proposal generation, features extracted from image mainly rely on histograms of gradient orientation (HOG [4,5]) or scale-invariant feature transform (SIFT [6]), and support vector machine (SVM [7]) or Adaptive Boosting [8] as the pedestrian classification methods. Those low-level features designed by hand-crafted receive good success. However, the feature extracted by this method is characterized by shallow features. The capacity of the feature is insufficient and is not suitable for complex pedestrian detection such as pedestrians in crowed scenes. The second approach is through the deep learning technology to achieve the goal of pedestrian detection, and have outperformed state-of-art performance on several pedestrian datasets. This method uses the convolutional neural network (CNN) to automatically extract the global and semantic features of the image, generate high-quality candidate boxes and classify each candidate. At present, the most classic method based on CNN is Faster R-CNN, which consists of two components: a region proposal network (RPN) to generate nearly cost-free region proposals, and then uses a Fast RCNNclassifier to detect pedestrians, which has shown leading accuracy on several multi-category benchmarks. After that researcher continue to develop new convolution neural network model, such as You Only Look Once (YOLO), Single Shot Multi Box Detector (SSD) and so on,

First, typical scenarios of pedestrian detection, such as automatic driving and intelligent surveillance, generally the size of pedestrian instances in small object image are less than 30x40 pixels. this make it difficult for the classifier to identify whether it is a pedestrian or a non-pedestrian.

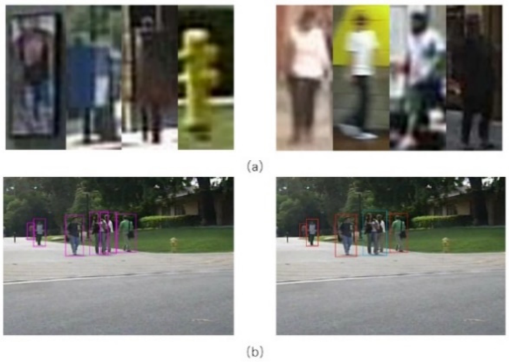

Figure 1. Challenges in pedestrian detection (a)hard negative samples of low resolution (b)Examples of pedestrians in crowed scenes. Left picture indicates ground truth, right picture denotes the prediction result.

pillar boxes, models in shopping windows and traffic signs, which have very similar apparent features with pedestrians.

[a] Corresponding author: 980552925@qq.com

without Sufficient features, detectors working with such low-resolution inputs are
unable to discriminate between them, and thus degrade the downstream classifier.

Second, Figure 1(b) is a common example in practical applications where the pedestrians stand close in a crowded scene. Under the circumstances, pedestrian detectors typically fail to locate each pedestrian and hence produce plenty of false positives owing to inaccurate localization.

In this paper, we propose to Faster RCNN-based pedestrian detection framework. Our work is also inspired by the work which aims at detecting small objects. We use VGG16 net as the baseline of our network. The pedestrian detection network extract low-level features from conv1, conv2 and conv3, and extract high-level from conv3 conv4 and conv5 of VGG16 simultaneously, the extracted feature maps are not sampled to same size. Region proposal network can locate more precise with the help of low-layers features while the detector is able to discriminate between pedestrian object and backgrounds with high-level semantic information. The evaluation results show that our network achieves the state-of-the-art performance on Caltech datasets.

2 Related Work

In the previous section, the paper has introduced many common methods for pedestrian detection. In this section we refer to some CNN based methods most related to our work.

Girshick et al. first introduced convolutional neural network framework for general object detection by Regions with CNN (RCNN). Compared to the previous method, RCNN introduced a classical detection framework: generation of region proposal and a CNN-based classifier, the popular object detection approach adopted the such pipeline, such as Fast/ Faster RCNN. R-CNN relies on region proposals generated by Selective Search, which first extracts about 2k region proposals and then classifies them with a pre-trained convolutional neural network.

3 Approach

In this section, we first describe detailed the proposed pedestrian detection framework and dataset we use.

3.1 Framework Architecture

Our framework is illustrated in Figure 2. As shown, our framework consists of three components. Due to the convolution operator with different number of convolution kernel and pooling operator will change the number of channel and the size of the features map, which trigger different layers has the feature maps are of different size and number of channels. In order to make each feature map have the same number of channels, we add two convolution layers after each feature map, while the multi-level maps are then subsampled to the same size as the fifth activation map.

3.2 Region Proposal Network (RPN) and Fast R-CNN

We build the RPN and Fast R-CNN using the same structure as proposed in [1]. For the RPN, unlike the original RPN in Faster RCNN, we adopt anchors of a single width to height aspect ratio of 0.41 owing to this aspect has been proved that it is the average aspect ratio in pedestrian detection task.

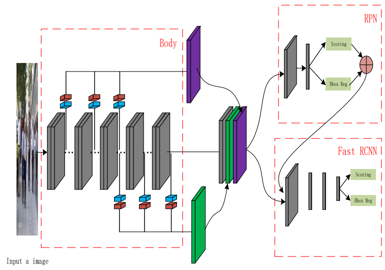

Figure 2. Pedestrian detection network with multi-levels features. Low-level features inhibit false positives of backgrounds and improve localization accuracy while high-level features can help detector discriminate hard positive samples and negative samples at low resolution.

3.2 Implementation Details

In our paper, our baseline detector is an implementation of Faster RCNN built on the popular deep learning framework Caffe, and the VGG16 architecture, all of which are available online. As is common practice, we use the pre-trained ImageNet model downloaded from the Caffe Model Zoo. Besides the two convolution layers are randomly initialized from a zero-mean Gaussian distribution with standard deviation of 0.01. For the RPN network, we assign positive anchor proposals that overlap the ground truth for more than 0.5 in intersection over union (IOU). Anchor proposals that overlap the ground truth for less than 0.3 in IOU are assigned as negative examples.

4 Experiments and results

In this section, we used a workstation with Intel i7-4790 3.6GHz CPU, 16GB memory, and TITAN X GPU with 12 GB of memory. We train and evaluate our proposed model on the well-known Caltech Pedestrian Dataset.

A. Caltech Pedestrian Dataset

The Caltech Pedestrian Dataset is the most popular and challenging dataset for pedestrian detection. It includes 350,000 annotations of 2300 unique pedestrians labeled in 250,000 frames which is about 10 hours at 30 Hz. [5] also analyzed the distribution of pedestrian pixel heights and histogram the heights of the 350,000 BBs using logarithmic sized bins. As for a further comment.

TableI Distribution of different scale on Caltech dataset

Scale	near	medium	far
Number of BBs	56000	241500	52500

We evaluate the results the dataset and compare with other methods including VJ, Faster RCNN, Deep Parts, MS-CNN, and RPN+BF. The performance metric used is log-average Miss Rate (MR) against False Positive Per Image (FPPI) evenly spaced in log-space in the range 10^{-2} to 10^{0}. We use the miss rate at FPPI= 10^{-1} as a common reference point to compare results. The experimental results are presented in Fig. 3. and Fig. 4.

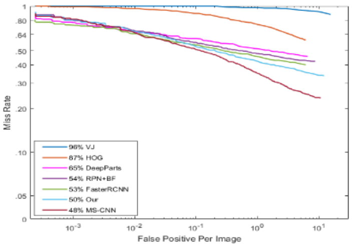

Figure 3. Performance comparisons of several well-known pedestrian detection methods in the near scale scenario on the Caltech benchmark.

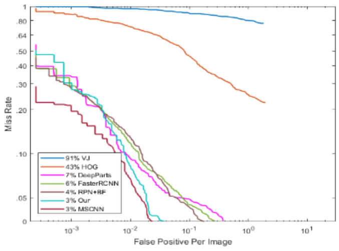

Figure 4. Performance comparisons of several well-known pedestrian detection methods in the medium scale scenario on the Caltech benchmark.

These results show that our approach achieves a 50% miss rate in near scale scenario and 3% miss rate in medium scale scenario, it achieved a slightly better performance than Faster RCNN and RPN+BF, which results indicate that small-size object detection performance can be improved by aggregated feature maps.

5 Conclusion

In this paper, we integrated low-level features and high-level features of the image into Faster R-CNN–based pedestrian detection system. Hierarchical activation map from multiple level has been proved to be useful, low-level features can improve localization precision while high-level features can help detectors discriminate hard positive samples. What's more, multiple level features increase the capacity of the feature extracted by CNN, and our quantitative experiments provide an acceptable miss rate in small-size object detection and precise localization. For future work, we want to explore the fusion of man-made features into convolutional neural network to make it even strong.

Reference

1. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Neural Information Processing Systems* (NIPS), (2015).
2. R. Girshick. Fast R-CNN. In International Conference on Computer Vision (ICCV), (2015).
3. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision,pages
4. 740–755. Springer, (2014)
5. Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), (2012).
6. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C] Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE: 886-893, (2005)
7. Y. Ke and R. Sukthankar. "PCA-SIFT: A more distinctive representation for local image descriptors". CVPR, Washington, DC, USA, 66-75, (2004).
8. S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. volume 0, pages 1–8, Los Alamitos, CA, USA. IEEE Computer Society.**1**, (2008)
9. Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst, **55**:119–139, (1997).