An Improved Location Model for Pedestrian Detection

Tengtao Zou\textsuperscript{1,}\textsuperscript{*}, Chen Cao\textsuperscript{1,b} and Shangming Yang\textsuperscript{1,c}

\textsuperscript{1}School of information and software engineering, University of Electronic Science and Technology of China

E-mail: \textsuperscript{*}ztjtzy@foxmail.com; \textsuperscript{b}1471421850@qq.com; \textsuperscript{c}minn003@163.com

Abstract. Pedestrian detection in complex scenes has always been a research difficulty in computer vision. The performance of current methods was seriously degraded when the pedestrian is occluded or the size of pedestrians are too small, etc. In this paper, we propose a novel approach based on location model for the detection of multiple pedestrians, which aims to improve the efficiency of algorithms in complex scenarios. In the model, a fully convolutional neural networks for the classification of pedestrian and non-pedestrian are trained to learn pedestrian features first. Then trained model are used to search for pedestrian regions, and the regions where pedestrian may be present will be activated and marked with boxes. Finally, we fine tune the boxes to overlap them with the ground truth more precisely. Compared with the current methods on two pedestrian datasets, experimental results demonstrate the comparable performance of our approach in term of miss rates.

Keywords: Pedestrian Detection, CNN, Location, Global Average Pooling.

1. Introduction

Pedestrian detection as the most popular and challenging problems in computer vision attracts lots of researchers. In the past ten years, a lot of good pedestrian detection algorithms \cite{1,2,3,4,5,6,7} have emerged. However, for CNN-based methods \cite{8,9,10,11,12}, such as Faster R-CNN, due to complexity of the scene, the performance of pedestrian detection is not very well. When the pedestrian is occluded or the size of the pedestrian is small, fully connected neural network are difficult to recognize these features so that the detector will neglect them. But in reality, these situations often occur. Thus, fast and accurate pedestrian detection in complex scenes is urgently needed.

In this paper, we propose a location based model for pedestrian detection. In order to obtain more efficient pedestrian detection, we improve the CAM \cite{13} to locate pedestrians. After pedestrian location, the image is separated into some regions, and a threshold can be set up to exclude some regions if their probabilities are lower than a certain threshold. The rest of region will be selected and marked with a box. To locate objects for convolutional neural networks more precisely, by replacing full-connected layers, we use full convolutional layers to be our training model.

The most related work to our algorithm is CAM \cite{13} and Faster-RCNN \cite{11}. CAM proposed by Zhou et.al, is a Channel-Wise Attention Mechanism, which is used to mine high-dimensional feature information of objects. Faster-RCNN as a generic object detector, contains RPN and Fast RCNN, in which RPN network was designed to generate a set of proposed regions instead of Selective Search and made use of Fast-RCNN for object recognition. In this paper, the advantages of both are exploited to construct pedestrian detection model.

The main contributions of this paper are listed as follows:(1) We propose an location based model...
using effective pedestrian region selection and improved pedestrian proposals. (2) The object location capability of CAM is improved by replacing full-connected layers. A full convolutional neural network is trained for object location. (3) Massive experimental results on two datasets demonstrate that the proposed algorithm significantly improves the detection rate pedestrian in complex scene.

2. The Proposal Model
In this section, we present the detailed implementation of our approach which includes three steps: pedestrian feature learning, region proposal and post processing. Firstly, we train a full convolutional neural network on the datasets that only include positive samples and negative samples (see Figure 1(a)), single pedestrian images and non-pedestrian images (see Figure 2). Secondly, the convolution kernels of trained convolutional neural network are transferred to the detection CNN for locating pedestrians. Then, the model will generate a pedestrian mask which displays the location of people. According to the pedestrian mask, we can use bounding boxes to mark the general position of pedestrians(see Figure 1(b)). At last, we fine tune the size and location of bounding boxes to make them more accurate and employ non-maximum suppression algorithm (NMS) [14] to delete redundant bounding boxes(see Figure 1(c)).

2.1. Pedestrian Feature Learning
To keep convolution nerve units learn pedestrian feature better, we should train a CNN-based network with two classifications. A CNN based on VGG [15] model is selected as our backbone network, and in the last convolutional layer, the network will produce 512 features maps.

After CNN network extracts features, the global average pooling is performed on the feature maps.
Then, the average result from global average pooling is used to indicate whether an input is pedestrian images (see Figure 1(a)). All parameters in network can be learned by minimizing the function:

\[ L = \frac{1}{n} \sum_{i=1}^{n} (-y_i \log(F(x_i)) - (1 - y_i) \log(1 - F(x_i))), \]

(1)

The datasets for training just include single pedestrian images and background images.

2.2. Region Proposal

In the previous section, we have completed the training of CNN-based network for classification. Then the convolution kernels will be transferred from classification network to detection network for pedestrian locating. An \( N \times M \) pixels image is input to the detection network to obtain deep convolutional features. the network will obtain 512 feature maps with the size of \( 1/16 \) of the input size. An average of all feature maps from the last convolutional layer will be computed to obtain the pedestrian mask (see Figure 1(b)). The pedestrian mask is with \( \frac{N}{16} \times \frac{M}{16} \) pixels, which respect to \( \frac{N}{16} \times \frac{M}{16} \) locations. Each element in the pedestrian mask indicates the probabilities of pedestrians appearing.

After the pedestrian mask are obtained, a threshold can be set up to exclude some locations if their probabilities are lower than a certain threshold. The rest of the regions are proposed. Then we map these proposed regions from these pedestrian masks to original image. Due to the proportion of pedestrian’s height and width is 2:1 generally, the proposed region is a bounding box whose ratio of height and weight is 2:1 approximately. Then a set of bounding boxes indicating position of pedestrians can be obtained.

2.3. Post Processing

To handle the variations of size and location of bounding boxes, we propose a multi-task regression model referring to RPN network. The multi-task regression model shares the convolutional features of proposed regions with detection model. We slide a small network over these convolutional features, which usually takes a 3*3 spatial window as the input of convolutional feature map. Each sliding window is mapped into lower-dimensional space. The features in low-dimensional space are fed into two sibling fully-connected layers: a box-adjustment layer and a box-score layer. The special model structure is shown in Fig 1(c).

In the box-adjustment layer, we use the parameterizations of four coordinates for bounding box fine-tuning, which are the same as the results in RPN. In the box-score layer, the IOU between predict box and ground truth is used to represent the score of bounding box, which is different from RPN. Therefore, the multi-task loss function is defined as following:

\[ L(p_i, t_i) = \frac{1}{n} \sum_{i=1}^{n} (l(p_i, p^*_i) + l(t_i, t^*_i)), \]

(2)

where Function \( l(x, y) \) is Smooth L1 Loss which is described in Fast-RCNN [10].

As training two regression networks, the parameters of the shared convolutional layer is fixed and only the parameters of the multi-task regression model is updated.

3. Experiment

To validate the effectiveness of the proposed method, extensive experiments on the INRIA person dataset [4] and the TUD-pedestrian dataset are conducted. The experimental results show that our method achieves the better performance by comparing with the generic detector (Faster RCNN) in terms of efficiency and accuracy. The MR(miss rate)-FPPI(false positive per image) and precision-recall (PR) are used as the evaluation metric.
Firstly, we compare the performance of the proposed algorithm and CAM on object location. Figure 3 shows a partial class activation maps obtained by CAM and our algorithm respectively. Obviously, our algorithm has better ability of occluded pedestrian location and small instance location than CAM.

From left to right, sub-figures in Figure 4 show the variation of miss rates of different methods [4, 10,11,12,16,17,18,19,20,21,22] on the two datasets. From this figure, we can see that our method obtains relatively good results among all methods in term of miss rate and precision. Figure 5 shows the results of pedestrian detection our method in the case of FPPI = 0.1 on INRIA person dataset and TUD-pedestrian dataset. The blue, green and red bounding boxes represent missing positives, false positives and true positives respectively.

![Figure 3: Some pedestrian activation maps obtained by CAM and our method respectively.](image)

![Figure 4: The variations of MR-FPPI of different methods on the INRIA and TUD-pedestrian dataset.](image)

![Figure 5: Some pedestrian detection results obtained our method on the INRIA person dataset and TUD-pedestrian dataset.](image)

### 4. Conclusions

In this paper, we propose an activation based region proposal model for pedestrian detection in complex scene. The model improves the structure of CAM by removing the full connected layer, which significantly improves the accuracy of pedestrian detection in complex scenes. Extensive experiments on two pedestrian datasets demonstrate the efficiency of the proposed model. As a future work, it is necessary to improve the regression model to make the bounding box more accurate.
ACKNOWLEDGMENTS
This research was supported in part by the National Natural Science Foundation of China (NSFC) under grants 61572107, the National Key Research and Development Program of China under grant 2017YFC1703905, and the Sichuan Science and Technology Programs under grant 2018GZ0192, 2018SZ0065.

Reference
[1] Shen X, Sui X, Pan K and Tao Y, 2016, Adaptive pedestrian tracking via patch-based features and spatial-temporal similarity measurement, Pattern Recognit, 163-173.
[2] Andriluka M, Roth S, and Schiele B, 2008, People tracking by detection and people detection by tracking, CVPR, 1-8.
[3] Viola P, Jones M and Snow D, 2003, Detecting pedestrians using patterns of motion and appearance, CVPR, 734-741.
[4] Dalal N and Triggs B, 2005, Histograms of oriented gradients for human detection, CVPR, 886-893.
[5] Gao W, Ai H, and Lao S, 2009, Adaptive contour features in oriented granular space for human detection and segmentation, CVPR, 1786-1793.
[6] Felzenszwalb P, Girshick R, McAllester D and Ramanan D, 2010, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell, 1627-1645.
[7] Hinterstoisser S, Lepetit V, Ilic S, Fua P and Navab N, 2010, Dominant orientation templates for real-time detection of texture-less objects, CVPR, 2257-2264.
[8] Yann L, Yoshua B and Geoffrey H, 2006, Deep learning. Nature, 436-436.
[9] Krizhevsky A, Sutskever I and Hinton G, 2012, ImageNet Classification with Deep Convolutional Neural Networks, 1097-1105.
[10] Girshick R, 2015, Fast r-cnn, CVPR, 1440-1448.
[11] Ren S, He K, Girshick R and Sun J, 2015, Faster r-cnn: towards realtime object detection with region proposal networks, NIPS, 91-99.
[12] Li X, Ye M, Liu Y and Zhu C, 2017, Memory-based pedestrian detection through sequence learning, ICME, 1129-1134.
[13] Zhou B, Khosla A, Lapedriza A and Oliva A, 2016, Learning Deep Features for Discriminative Localization, CVPR, 2921-2929.
[14] Neubeck A and Van L, 2006, Efficient non-maximum suppression, ICCV, 850-855.
[15] Simonyan K and Zisserman A, 2015, Very deep convolutional networks for large-scale image recognition, ICLR.
[16] Li X, Ye M, Liu Y, Zhang F, Liu D and Tang S, 2017, Accurate object detection using memory-based models in surveillance scenes, Pattern Recognition, 73-84.
[17] Viola P and Jones M, 2004, Robust Real-time Object Detection, IJCV.
[18] Dollr P, Tu Z, Perona P and Belongie S, 2009, Integral channel features, in: British Machine Vision Conference, 91-98.
[19] Dollr P, Belongie S and Perona P, 2009, The fastest pedestrian detector in the west, BMVC.
[20] Piotr D and Pietro P, 2014, Fast Feature Pyramids for Object Detection, IEEE Trans. Pattern Anal. Mach. Intell, 1532-1545.
[21] Nam W, Dollar P and Han J, 2014, Local decorrelation for improved pedestrian detection, NIPS, 424-432.
[22] Zhang L, Lin L, Liang X and He K, 2016, Is Faster R-CNN Doing Well for Pedestrian Detection, ECCV, 443-457.