Convolutional Neural Networks with Neural Cascade Classifier for Pedestrian Detection

Bei Tong, Bin Fan and Fuchao Wu

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{bei.tong,bfan,fcwu}@nlpr.ia.ac.cn

Abstract. The combination of traditional methods (e.g., ACF) and Convolutional Neural Networks (CNNs) has achieved great success in pedestrian detection. Despite effectiveness, design of this method is intricate. In this paper, we present an end-to-end network based on Faster R-CNN and neural cascade classifier for pedestrian detection. Different from Faster R-CNN that only makes use of the last convolutional layer, we utilize features from multiple layers and feed them to a neural cascade classifier. Such an architecture favors more low-level features and implements a hard negative mining process in the network. Both of these two factors are important in pedestrian detection. The neural cascade classifier is jointly trained with the Faster R-CNN in our unifying network. The proposed network achieves comparable performance to the state-of-the-art on Caltech pedestrian dataset with a more concise framework and faster processing speed. Meanwhile, the detection result obtained by our method is tighter and more accurate.

Keywords: Convolutional Neural Network, Cascade classifier, Faster R-CNN, Pedestrian detection

1 Introduction

Object detection is an enduring topic in the field of computer vision. As a typical issue of object detection, pedestrian detection attracts increasing attention in the field of surveillance, autonomous driving and robotics applications. Since the robust real-time face detection method [30] was proposed, it was widely applied to pedestrian detection. Histogram of Oriented Gradient (HOG) [10] accelerated the development of pedestrian detection and led to the formation of the framework of features and classifier. Dollár et al. [14][13][12][11][20] proposed ten channel features and a cascade AdaBoost classifier based on [30] and Felzenszwalb et al. [16][18][17] presented the usage of Deformable Part-based Model (DPM) in pedestrian detection. Both of them had a great influence on the later methods.

As a fundamental component of pedestrian detection, feature extraction is vital to the subsequent processes. Various hand-crafted features had been proposed such as CSS [31], InformedHarr [34], Motion [23], Cross Channel feature

[33], Checkboard filters [36]. Researchers try to design more discriminative features since it has been proven that the combination of more features can bring better performance to a certain extent. However, more features also require more computations. The performance improvement is achieved with a sacrifice of the model's efficiency. What's more, although this kind of methods is effective, hand-crafted features are difficult to design and may be too task-specific to extend on more diverse datasets and general object detection tasks.

Recently, using CNNs to automatically learn features has become a tendency. VGG16 [26] was applied to extract features and a cascade AdaBoost classifier was trained based on these features [32][7]. Their good performance testified that CNNs have a strong power of extracting general and representative features without the need of human interference. However, these methods were always based on rectangle window of a fixed size, so they had to firstly get proposals provided by other methods such as ACF [11], stixel [3][2], Edge Boxes [37], BING [9], selective search [29], Objectness [1] and CPMC [8]. Moreover, their methods were not end-to-end and several stages of processes had to be gone through before giving the final results. Despite they had achieved good performance, sophisticated operations limit their practical use and may be time-consuming as well.

In this paper, we propose an end-to-end neural network based on Faster R-CNN [24] and neural cascade classifier for pedestrian detection. Mainly derived from the architecture of Faster R-CNN, our network is free of external proposal extraction methods and hand-crafted features. Moreover, motivated by the idea that low-level feature maps carry local information of the image [7], we utilize features from multiple convolutional layers rather than the last one to incorporate more information. These features are fed to a neural cascade classifier for hard negative mining and pedestrian detection. The neural cascade classifier consists of multiple softmax classifiers and helps to filter out negative samples in each classification stage. By integrating Faster R-CNN with neural cascade classifier, we build a unifying neural network whose inputs are images and outputs are the corresponding bounding boxes (bboxes) with the confidences of including a person. The proposed network makes a step closer to the real-time pedestrian detection since it can process an image in about 0.7 seconds. What's more, it achieves comparable performance to the state-of-the-art on Caltech pedestrian dataset with a more concise framework and can be extended to diverse object detection tasks.

Overall, the contributions of this paper and the merits of the proposed network can be summarized as follows:

1) We propose an end-to-end neural network based on the Faster R-CNN and cascade neural classifier. The network does not resort to any hand-crafted features. All features are learnt by the network automatically.

2) We utilize both low-level and high-level features and build a neural cascade classifier for hard negative mining. The cascade mechanism not only boosts the classification performance of the network, but also accelerates the processing procedure by quickly rejecting the majority of negatives.

3) The network performs well on the Caltech dataset with the new anno-

tations [35] and runs very fast. Besides, our network is more elegant than the frameworks proposed by the previous methods.

The remainder of this paper is organized as follows. Firstly, we give a review of related works about pedestrian detection in Sec.2. In Sec.3, we introduce our end-to-end neural network and implementation details. Experimental results are shown and discussed in Sec.4. Finally, we conclude this paper in Sec.5.

2 Related work

Although excellent performance has been achieved on the Caltech reasonable subset, pedestrian detection still has a long way to go for the following reasons. 1) The false positive rate and false negative rate of the detection results are not satisfactory, let alone the speed of the model. 2) The original evaluation protocol [11] is not enough to describe the model's performance and there is still a large gap between the state-of-the-art results and the human baselines according to [35]. 3) Occlusion is impossible to neglect since nearly 70% of the pedestrians captured in street scenes are occluded in at least one video frame according to [15], while many state-of-the-art methods do not consider the occlusion problem.

Recent pedestrian detection methods can be divided into three categories: hand-crafted feature based methods, CNNs based methods and combination (i.e., mixture of traditional methods and CNNs) based methods. The first category contains two mainstream branches, decision forest [11][20][34][36][4] and DPM variants [16][18][17] according to [5]. Both are based on hand-crafted features and traditional classifiers such as AdaBoost, SVM, etc. Most of these methods need to construct multilayer pyramid models and test by sliding window. Although the training and testing speeds of these methods are considerable at an early stage, they become slower when more features are used for improving performance. The second category becomes popular with the upsurge of deep learning in pedestrian detection. In the beginning, researchers tend to design and train their own networks for a certain task. Sermanet et al. [25] utilized an unsupervised method based on convolutional sparse coding to pre-train the filters at each stage and then trained a pedestrian detection model with multi-stage features. Wang et al. [21][19] designed their unique network structure as well. These shallow networks do make effects, but the performance is not very well. With the advent of VGG, GoogLeNet and other very deep CNNs, researchers find that for a certain task, fine-tuning these CNNs pre-trained with massive general object categories brings much better performance. Thus in the later period, most CNN based methods fine-tune these famous CNNs with a specific task. The third category is born because the combination of the previous two families [7][35] makes sense. It can integrate their advantages and be effective in detection and classification, but the tedious processes limit its possibility of application.

Overall, it is hard to judge which kind of method is obviously better than others since the state-of-the-art methods come from all three categories. How-

ever, there is no doubt that deep learning methods used in pedestrian detection form a tendency. Unfortunately, most of exiting methods are complicated and inelegant. Tian et al. [27] trained 45 different part models based on the proposals provided by LDCF [20] and selected 6 models for testing in order to save time. Each model was an independent VGG16 network and a SVM classifier was applied for combining these models to give the final result. The framework was time-consuming and not an end-to-end network. In [6][35], traditional methods were firstly applied to output a moderate good result and the best results were achieved after the binary classification via VGG16. For these methods, the traditional part requires much time to grasp, let alone the additional cost caused by VGG16. By contrast, the method proposed in this paper is more concise and its result is comparable to the state-of-the-art results. No hand-crafted features are introduced into the model and a single network is learnt automatically. Namely, the network is end-to-end without other operations.

3 Models

3.1 Datasets

Our model is trained on the Caltech10× pedestrian dataset relabeled by [35] since the old annotations have many wrong labels or labels shifting away from the real objects. Fig. 1 shows two samples of the Caltech training dataset. The red bboxes are the old annotations and the green ones are the new annotations. It is obvious that the new annotations are more accurate than the old ones. However, the new labelled annotations also have some problems, such as missing or shifting labels, which can be seen from Fig. 1(b). More details can refer to [35]. As our model is trained image by image, namely batch size of 1, wrong labels in one image may hurt the network's learning process and lead to a suboptimal converge path. Thus, the new and more accurate annotations are chosen for training and evaluation.

3.2 Architecture

CNNs have great potential in feature extraction, which can be seen in [25][21][19] [28][32][27][22]. To avoid sophisticated hand-crafted features, our model utilizes convolutional layers to obtain features as well. The proposed network integrates a pedestrian proposal network and a neural cascade classifier in a unified framework. The pedestrian proposal network aims to predict the location of pedestrian and provides the confidence of the predicted rectangle boxes simultaneously. The neural cascade classifier attempts to give more accurate results through classification of several stages and enables the network to filter out negatives in an elegant way. Both structures share all the convolutional layers, which could obtain general characteristics and avoid too many parameters. The network takes the whole image as the input, and directly outputs the detection results with corresponding bboxes by no means of any external operations.

(a) (b)

Fig. 1. Samples of Caltech train set. Red bboxes present the old annotations and the green ones are the new annotations. The left image shows that the old annotations have several pixels offset away from the objects and some missing bboxes, which may result in regarding the true positive detection bboxes as false positive if the Intersection Over Union (IOU) threshold is set high. The right image indicates that the new annotations still have some problems of the missing and shifting labels and need further revision. However, in the current stage, we choose relatively accurate dataset (i.e., new annotations provided by [35]) for training and testing.

The basic architecture of our network is shown in Fig. 2. It consists of two parts: proposal extraction network (i.e., Part1) and neural cascade network (i.e., Part2). Since the best performance is achieved with a cascade classifier of two stages that is demonstrated in Sec.4, the cascade classifier we refer to contains two-stage classification if not specified.

The input of the network is the whole image whose size is 800×600, which is resized from the Caltech train image whose size is 640×480. The resize operation enlarges the size of pedestrian and meanwhile maintains the aspect ratio of the whole image, which enables the network to deal with larger pedestrians since small size pedestrians are very difficult to detect. Then, several convolutional layers are applied to extract features, which are the same as Faster R-CNN except the absence of *pool3* and *pool4*. This is because that pedestrian in Caltech dataset is much smaller than the object in PASCAL VOC. To keep the pedestrian area not too small and guarantee adequate foreground anchors for *Region Proposal Network* (RPN) training, *pool3* and *pool4* operations are removed. The sizes of feature maps are listed in Table 1. After the shared convolutional layers, the proposal extraction network is firstly applied to predict the location of the pedestrian and meanwhile outputs two scores of each bbox. The neural cascade network is then used to filter out the negatives from the top $2,000$ bboxes sorted by scores provided by the Part1. It can be divided into two stages and a bbox selection mechanism is employed between the two stages. The inputs of the first stage are the predicted bboxes and the corresponding region features from conv4_3. Then the combination of 512 dimensional fully connection layer and binary softmax classifier is adopted to eliminate negative bboxes. Subsequently,

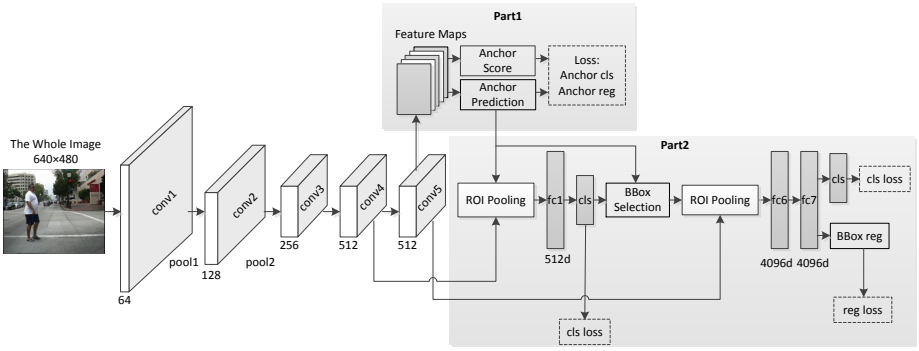


Fig. 2. Architecture of our model with two-stage classification. It can be divided into two parts, proposal extraction network and neural cascade network. The parameters in all the convolutional layers are shared by the two parts. The first part aims to predict the bbox's location and the second part is used to classify the proposals provided by the first one.

the bboxes regarded as foreground bboxes by the first stage are fed to the next stage through a bbox selection mechanism. In the network training phase, random sampling in predicted bboxes is employed when the number of foreground bboxes is less than 64. The second stage is followed with two $4,096$ dimensional fully connection layers and one binary softmax classifier. Its inputs are the region features from conv5_3 and the outputs are the detection results.

Table 1. The size of the feature map of each convolutional layer. The original image is resized before feeding to the network. Only the top two pooling operations of VGG16 are preserved since the RPN can gain more foreground bboxes for training.

Layer	L0	L1	L2	L3	L4	L5
Name	Original Image	VGG16				
		conv1	conv2	conv3	conv4	conv5
Channels	3	64	128	256	512	512
Size(width×height)	640×480	800×600	400×300	200×150	200×150	200×150

The proposed neural cascade structure utilizes multilayer features and forms a strong classifier. It can quickly reject the majority of negative bboxes in the early stages and relief the burden of computation for the following stages, which is much helpful for accelerating processing speed and improving the output accuracy.

3.3 Aspect Ratio and Scale

Traditional proposal extraction methods such as ACF [11] use one aspect ratio of all detection bboxes if one model is employed. Several models have to be trained

if different aspect ratio bboxes are need. Thanks to the characteristic of the Caltech test set whose pedestrian's aspect ratio is around 0.41, one aspect ratio has little effect on the model's final performance when evaluating on this dataset because its groundtruth bboxes are resized to guarantee a constant aspect ratio.

However, it is unreasonable to restrict the aspect ratio to a constant value in practical use due to the fact that the aspect ratios of pedestrians captured in street scenes distribute in a wide range. As a consequence, the proposed network should have the ability to deal with objects of different aspect ratios. There are usually two ways to solve this problem. One is resizing all bboxes to a fixed size and the other is building several models of different aspect ratios. RPN can deal with several different aspect ratios with one model by producing anchors when different aspect ratios and scales are set. The proposal extraction network of our model is based on RPN.

In RPN, more different aspect ratios and scales mean more different anchors, and theoretically should have more bboxes with high IOU values to the groundtruth. However, there is a trade-off between the number of anchors and the computation time. As a result, several relatively better aspect ratios and scales are chosen for training and testing according to the experimental results. Table 2 shows the results of using different configurations of aspect ratios and scales when three aspect ratios and scales have to be selected. The third group of configurations are chosen for an overall high performance.

Table 2. Results of configurations of different aspect ratios and scales. Three sets of one thousand images are randomly sampled from Caltech training dataset since it is time-consuming to take all training data into consideration. Aspect ratio and scale are generated randomly in a reasonable range. Only 9 different anchors (3×3) are used to compromise between time and precision. Four indicators including total_gt (i.e., total number of groundtruth bboxes), total_iou_fg (i.e., total number of bboxes whose IOU value is larger than 0.7 with any groundtruth bbox), total_fg (i.e., total number of bboxes regarded as foreground bbox) and zero_num (i.e., total number of images whose IOU value of generated bboxes with any groundtruth bbox is smaller than 0.7) are shown. The table lists only top ten results among $1,200$ different random configurations sorted by total_iou_fg. A configuration should result in high total_iou_fg, total_fg and low zero_num with a smaller total_gt.

	aspect ratio	scale	total_gt	total_iou_fg	total_fg	zero_num
1	$2.2, 2.3, 2.4$	$3.7, 7, 4.3$	$2,049$	$36,713$	$71,874$	436
2	$3, 2.3, 2.2$	$3.4, 3.6, 6.8$	$2,077$	$36,444$	$71,903$	448
3	$2.4, 2.9, 2.8$	$4.4, 2.9, 10.6$	$2,132$	$34,896$	$67,054$	239
4	$1.7, 2.9, 2.7$	$2.8, 3.8, 3.7$	$2,132$	$34,299$	$66,739$	283
5	$1.9, 2.8, 2.2$	$6.8, 3.8, 4.3$	$2,077$	$33,957$	$70,660$	469
6	$2.8, 1.9, 2.3$	$3.4, 4.2, 8.4$	$2,049$	$33,444$	$65,826$	414
7	$2.7, 2.9, 1.8$	$3, 7.1, 3.5$	$2,049$	$32,868$	$62,700$	277
8	$2.1, 2.6, 2.9$	$5.1, 3.7, 6$	$2,132$	$32,328$	$69,107$	406
9	$2.2, 1.5, 2.9$	$4, 6.6, 3.9$	$2,077$	$32,277$	$66,982$	456
10	$2.8, 3.6, 2.5$	$3.2, 6.5, 3.4$	$2,077$	$32,057$	$69,834$	386

3.4 Optimization

Since the foreground bboxes of the top $2,000$ bboxes predicted by the Part1 are limited at most 32, ordinary softmax loss is no longer applicable. All bboxes will be classified as negatives by the fully connection layers because of the huge difference between the number of background and foreground bboxes when the simple softmax loss is applied in back propagation. As a consequence, different loss weights are employed to negatives and positives. According to the approximate ratio of background and foreground bboxes, the loss is multiplied by 0.5 if the bbox is negative and 19 if otherwise.

4 Experiments

The Caltech train set and reasonable test set are employed for training and evaluation respectively. Due to the reasons explained in Sec.3.1, the new annotations of Caltech10× dataset [35] are used to train our network with the initial VGG16 weights pre-trained on the ImageNet dataset. Since far or occluded person is not considered in the training data, the total number of qualified images for training is only $15,678$, including mirrored images. Reasonable test set is a subset of the test dataset in which the pedestrian's height is larger than 49 pixels and the percentage of visual part is larger than 65%. It is the most frequently used test set and evaluation on it is considered more representative than evaluated on the whole test set.

First of all, we give an overview of the training and testing processes. Fig. 3 presents the pipelines of training and testing processes respectively. The main processes are similar and the input of both is the whole image and the output is the detection result. Besides using the groundtruth bboxes to compute the loss for back propagation and training RPN, another difference between training and testing processes is the bbox selection after each stage. In the training process, there are at least 64 bboxes to be judged in each stage, while in the testing process, any bbox which is classified as negative bbox in the previous stage will not appear in the next stage. Both processes are end-to-end and do not need any additional operations.

To verify how many stages make the best performance in terms of precision and time, a series of experiments have been conducted with one to four stages. The input batch size of the cascade with only one stage is restricted to 128 since one stage structure can not filter out part of bboxes in advance and too many bboxes fed to the next two $4,096$ dimensional fully connection will sharply increase the burden of computation. The other structures with more than one stage do not need this additional operation because they can reject part of bboxes by a smaller fully connection layer beforehand. This results in the training time of one stage structure is the least among all the tested structures, which is 15 hours for training with $10,000$ iterations. The corresponding time cost by structures with two to four stages are 22 hours, 19 hours and 21 hours respectively. Generally, more stages guarantee less time to train since the fewer bboxes need to be considered by the larger fully connection behind. However, at least 64 bboxes are

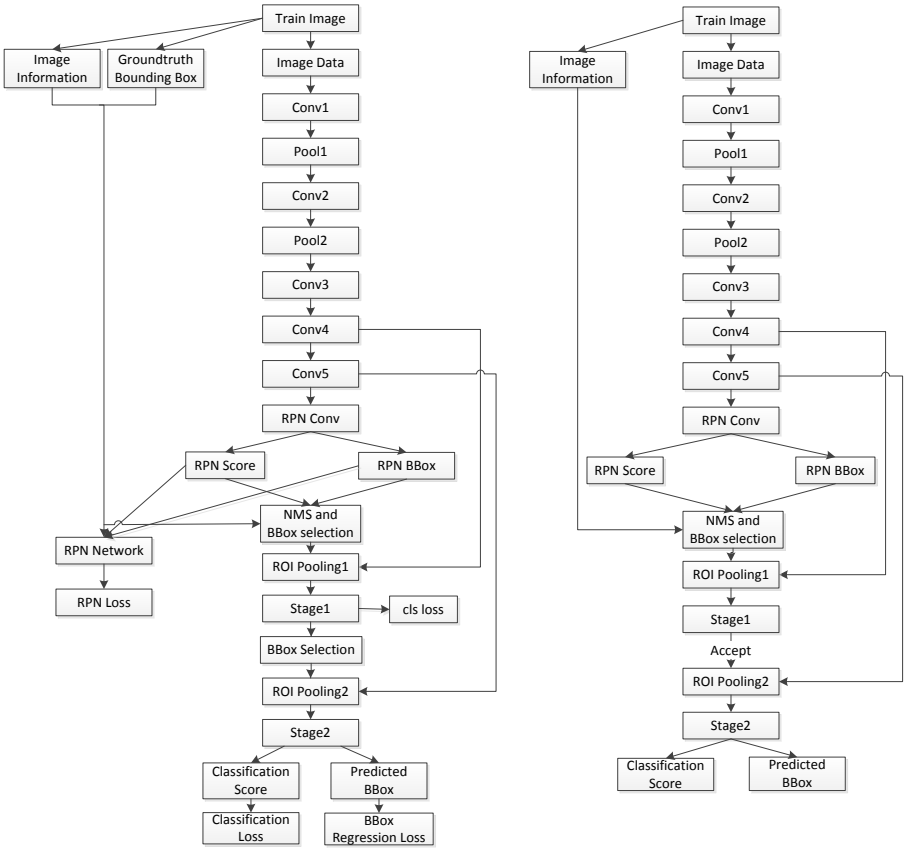


Fig. 3. Flow chart of the training (left) and testing (right) procedures. Conv1 to Conv5 present the five large convolutional layers of VGG16. RPN is used to extract proposals. More details can refer to [26][24].

remained to guarantee the minimal batch size for the fully connection of each stage. This is the reason why the training time of three-stage structure is less than that of the four-stage structure.

Table 3 gives the miss rate of our network on the test set with varying number of classification stages. All parameters are the same except the stages of classification. The experimental results show that cascade softmax classifer works well in eliminating negative bboxes and two-stage classification already performs well. The output of the network is evaluated by the Caltech evaluation program [11] after Non-Maximum Suppression (NMS) operations with an IOU value of 0.65, which is the same as the output of ACF [11].

It can be known from Table 3 that classifier with more stages does not guarantee better performance. We think it is due to the small size of train set since the network tends to overfit with more fully connection layers. In addition, more

Table 3. Evaluation results of classifier with different stages. The miss rate is given by the Caltech MATLAB evaluation program [11] evaluated on the Caltech reasonable test set with the new and more accurate annotations [35]. The *Time* is the approximate average time of testing on one image and the *BBox* is the number of bboxes outputted by the model with the lowest miss rate among the $80,000$ iterations detected on the test set after NMS operation.

Train Iterations	One Stage		Two Stages		Three Stages		Four Stages	
	Miss Rate	Time	Miss Rate	Time	Miss Rate	Time	Miss Rate	Time
$10,000$	23.56%	$0.75s$	13.78%	$1.0s$	13.49%	$0.68s$	14.19%	$0.7s$
$20,000$	21.53%	$0.75s$	11.72%	$0.7s$	11.35%	$0.88s$	13.85%	$0.9s$
$30,000$	20.71%	$0.81s$	11.28%	$0.9s$	12.59%	$0.72s$	14.79%	$0.69s$
$40,000$	19.88%	$0.76s$	11.44%	$0.7s$	12.45%	$0.76s$	14.88%	$0.88s$
$50,000$	19.75%	$0.71s$	11.35%	$1.2s$	12.33%	$0.7s$	14.65%	$0.71s$
$60,000$	19.58%	$0.74s$	11.06%	$0.7s$	12.52%	$0.69s$	15.03%	$1.21s$
$70,000$	19.55%	$1.5s$	11.10%	$0.71s$	12.29%	$0.69s$	14.88%	$0.7s$
$80,000$	19.58%	$0.85s$	11.15%	$0.68s$	12.37%	$1.1s$	14.98%	$0.68s$
BBox	$47,758$		$15,927$		$5,755$		$5,356$	

stages mean a stronger classifier, which may cause some bboxes of low score such as cyclist, occluded person to be rejected by the classifier. The network can converge to a good result around $20,000$ iterations. Although more iterations could bring better training performance, it leads to less testing accuracy due to overfitting. The network could potentially perform better if there are more different training images since the Caltech train set is sampled from a video with around $2,300$ unique pedestrians and its diversity is far from rich. For instance, the low confidence of cyclist may be due to the biased train set which contains few cyclists.

Since all the experiments are carried out on the server cluster, the statistics of testing time are not very stable due to the discrepancy in load capacity. However, the majority of testing times of classifier with different stages are close. Nearly 0.7 seconds are cost for testing each image and 11.06% miss rate is achieved on the test set. For a compromise of performance and time, two-stage classification is applied and the best result we achieved is used for comparing with the state-of-the-art methods.

Fig. 4 compares our model with recent state-of-the-art methods evaluated on the same test set. Average miss rates over the FPPI range of $[10^{-2}, 10^{0}]$ (MR_{-2}) are shown. In the brackets, we also show the average miss rates over the range of $[10^{-4}, 10^{0}]$ (MR_{-4}). Fig. 4(a) presents the results with IOU being 0.5, which means a detection bbox is regarded as a true positive when its IOU with any groundtruth in the same image is larger than 0.5. It can be seen that our model achieves a second lower MR_{-2} as 11.06%, which is 5.44%, 3.41%, 0.32% lower than those of TA-CNN, Checkerboards and DeepParts respectively. Similar results could be observed for MR_{-4}. Fig. 4(b) shows the results of IOU being 0.65. The higher the IOU threshold is, the more accurate and tighter the detection results are. From this figure, we can find that our model significantly outperforms

other methods, with a decrease of 18.62%, 14.55% and 11.14% compared to TA-CNN, Checkerboards and CompACT-Deep respectively. Average miss rates over the FPPI range of $[10^{-2}, 10^{0}]$ with other IOU settings are listed in Table 4. This result demonstrates that the detection results of our model are tighter and closer to the groundtruth.

Table 4. Average miss rates over the FPPI range of $[10^{-2}, 10^{0}]$ with different IOUs.

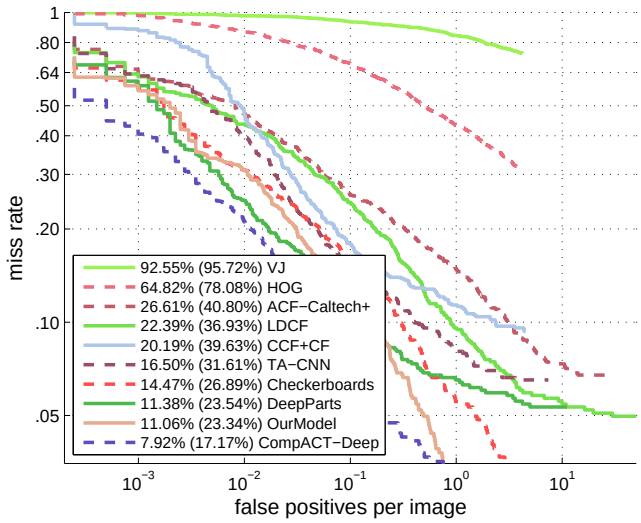
IOU	TA-CNN	Checkerboards	DeepParts	CompACT-Deep	OurModel
0.5	16.50%	14.47%	11.38%	**7.92**%	11.06%
0.55	20.59%	17.22%	15.18%	**11.48**%	11.86%
0.6	26.02%	22.91%	22.55%	16.66%	**12.75**%
0.65	34.27%	30.20%	34.90%	26.79%	**15.65**%
0.7	45.01%	42.55%	53.02%	39.16%	**19.21**%
0.75	61.33%	58.30%	68.54%	59.14%	**27.16**%
0.8	80.11%	76.01%	80.64%	76.58%	**42.05**%
0.85	91.33%	88.63%	91.31%	89.23%	**64.13**%
0.9	97.28%	96.14%	97.56%	96.36%	**86.17**%

The number of bboxes detected on the test set of different methods are summarized in Table 5. It can be seen that our model has the fewest bboxes with an excellent performance. What's more, the three-stage structure can reduce the number to about one third of that of the two-stage structure with a 0.29% increase of miss rate over the FPPI range of $[10^{-2}, 10^{0}]$. It indicates that our model produces smaller number of false positives, which is extremely important for practical use since too many false positives will increase the processor's burden.

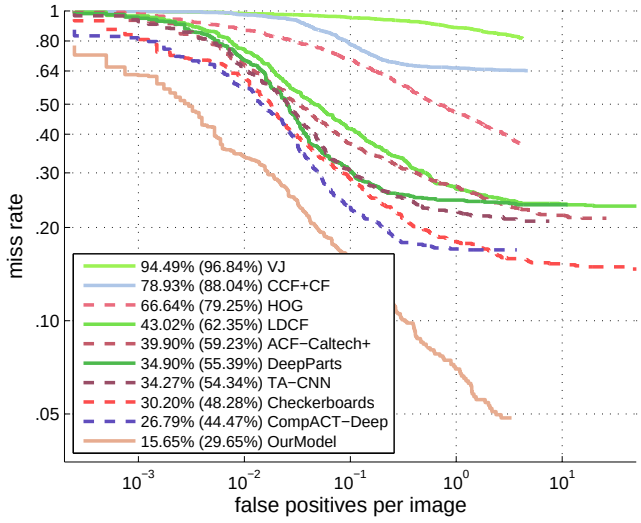
Since training the model is time consuming, limited by our computational resources, our hyperparameter tuning strategy is relatively coarse-grained. More fine-grained tuning should bring better performance. Although our model's performance does not exceed [35][6], it is better than them if no additional VGG16 is applied to further classify the results of them. Overall, the model has following merits. 1) Our model is more concise and convenient for training and testing instead of dividing the train and test processes into several parts. 2) Although it is hard to compare all models' runtime, our model is at the leading level considering both performance and time according to the summarization of [7][6][5][11]. 3) The detection results of our model are tighter which can be seen from Table 4. 4) With a tight IOU threshold, our model achieves the best results.

5 Conclusion

In this paper, we have proposed an end-to-end neural network based on Faster R-CNN and neural cascade classifier. It uses multiple convolutional layers to learn rich features, which are shared by the RPN for predicting the locations of bboxes



(a)



(b)

Fig. 4. Performance comparison to the state-of-the-art methods. (a) and (b) draw the ROC curves with IOU being 0.5 and 0.65 respectively. The results show that our model performs well and is comparable to the state-of-the-art.

Table 5. The number of bboxes of different methods.

Method	VJ	HOG	ACF-Caltech+	LDCF	CCF+CF
BBox	190,867	33,508	106,855	225,702	19,706
Method	TA-CNN	Checkerboards	DeepParts	CompACT-Deep	OurModel
BBox	31,676	1,487,711	46,684	16,337	15,927

and the muti-stage softmax cascade classifier used for pedestrian classification. The model is concise and achieves comparable miss rate to the state-of-the-art with more accurate detections and being faster to process. The network can run on Tesla K80 by around $0.7s$ for each image. Future work will be on improving the model's capability to deal with occlusion since most pedestrians captured in the street are occluded.

6 Acknowledgement

This work was supported in part by the Projects of the National Natural Science Foundation of China (Grant No. 61375043, 61403375, 61272394) and the Beijing Natural Science Foundation (Grant No. 4142057).

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11), 2189–2202 (2012)
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Fast stixel computation for fast pedestrian detection. In: Computer Vision–ECCV 2012. Workshops and Demonstrations. pp. 11–20. Springer (2012)
3. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2903–2910. IEEE (2012)
4. Benenson, R., Mathias, M., Tuytelaars, T., Gool, L.: Seeking the strongest rigid detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3666–3673 (2013)
5. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Computer Vision-ECCV 2014 Workshops. pp. 613–627. Springer (2014)
6. Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3361–3369 (2015)
7. Cao, J., Pang, Y., Li, X.: Learning multilayer channel features for pedestrian detection. arXiv preprint arXiv:1603.00124 (2016)
8. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric min-cuts. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(7), 1312–1328 (2012)

9. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: Bing: Binarized normed gradients for objectness estimation at 300fps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3286–3293 (2014)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
11. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(8), 1532–1545 (2014)
12. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. In: Computer Vision–ECCV 2012, pp. 645–659. Springer (2012)
13. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC. vol. 2, p. 7. Citeseer (2010)
14. Dollár, P., Tu, Z., Perona, P., Belongie, S.J.: Integral channel features. In: BMVC. pp. 1–11. British Machine Vision Association (2009)
15. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(4), 743–761 (2012)
16. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
17. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. pp. 2241–2248. IEEE (2010)
18. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(9), 1627–1645 (2010)
19. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 899–906 (2014)
20. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems. pp. 424–432 (2014)
21. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2056–2063 (2013)
22. Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.C., et al.: Deepid-net: Deformable deep convolutional neural networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2403–2412 (2015)
23. Park, D., Zitnick, C., Ramanan, D., Dollár, P.: Exploring weak stabilization for motion feature extraction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2882–2889 (2013)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015)
25. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3626–3633 (2013)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

27. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1904–1912 (2015)
28. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5079–5087 (2015)
29. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2), 154–171 (2013)
30. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision 57(2), 137–154 (2004)
31. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. pp. 1030–1037. IEEE (2010)
32. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 82–90 (2015)
33. Yang, Y., Wang, Z., Wu, F.: Exploring prior knowledge for pedestrian detection. In: Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015. pp. 176.1–176.12 (2015)
34. Zhang, S., Bauckhage, C., Cremers, A.: Informed haar-like features improve pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 947–954 (2014)
35. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? arXiv preprint arXiv:1602.01237 (2016)
36. Zhang, S., Benenson, R., Schiele, B.: Filtered feature channels for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2015)
37. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014, pp. 391–405. Springer (2014)