Taking a Deeper Look at Pedestrians

Jan Hosang, Mohamed Omran, Rodrigo Benenson, Bernt Schiele
MPI for Informatics, Saarbrücken, Germany

In this paper we study the use of convolutional neural networks (convnets) for the task of pedestrian detection. Despite their recent diverse successes, convnets historically underperform compared to other pedestrian detectors. Current top performing methods are all based on decision trees learned via Adaboost. In contrast to previous work on convnets for pedestrian detection [2, 3, 4, 5, 6, 7], we deliberately omit explicitly modelling the problem into the network (e.g. parts or occlusions) and show that we can reach competitive performance without bells and whistles. We report experiments analysing small and big convnets, their architectural choices, parameters, and the influence of different training data, including more frames from the Caltech training videos and pre-training on surrogate tasks.

We present the best convnet detector results on the Caltech and KITTI dataset. On Caltech our convnets reach top performance both for the Caltech1x and Caltech10x training setup. Using additional data at training time our strongest convnet model is competitive even to detectors that use additional data (optical flow) at test time.

A continuum of convnet architectures We experiment with convnet architectures that contain between 10^5 and 10^7 parameters. On the low end of the spectrum, we start from an architecture that is designed to solve the CIFAR-10 classification problem (CifarNet), while the biggest network we experiment with, known as AlexNet, is designed to solve the ILSVRC2012 classification problem. All networks we experiment with are generic in the sense that they do not contain components to specifically model pedestrians, they only consist of standard convolutions, pooling, contrast normalization, and fully connected layers.

As visualized in figure 1, the CifarNet consists of three convolutional layers followed by a fully connected layer. Figure 2 depicts the AlexNet with five convolutional layers and three fully connected layers.

Best convnet results on Caltech with a small, vanilla network Our experiments show that the small CifarNet is able to improve over all previous convnet pedestrian detectors. (This is even true if we use the same proposals; better proposals improve results further.) The CifarNet training is sensitive to parameters such as receptive field size, details of sampling training data, and the specific type of layers used, but when tuned properly it obsoletes previously published hand designed pedestrian specific convnets.

Although the AlexNet has two orders of magnitude more parameters, it is only two percent points log-average miss-rate worse than the CifarNet when trained on Caltech1x only (CifarNet 30.7% MR, AlexNet 32.4% MR, see figure 3).

The benefit of additional training data The training data of the Caltech dataset consists of videos and typically every 30th frame is used for extracting training data (named Caltech1x). By sampling training frames more densely, it is possible to obtain a lot more – albeit correlated – training data. We extend the training data to every 3rd frame and call this training set Caltech10x. We also experiment with pretraining convnets on ImageNet, which has proved useful for object detection [1].

As shown in figure 3, the CifarNet improves from 30.7% log-average miss-rate to 28.4% by extending the training data to Caltech10x. AlexNet improves by a larger margin from 32.4% to 27.5% MR. Although the data in Caltech10x is highly correlated to Caltech1x, this experiment shows that Caltech10x still contains non-negligible additional information. The experiment also suggests that the CifarNet lacks capacity to fully benefit from the larger training set.

Top performance with ImageNet pretraining Pretraining of the AlexNet on ImageNet improves performance to 23.3% log-average miss-rate, which is better than the previously best single frame detector LDCF with 24.8%. The performance is only 2 percent points short of the performance of best approaches that use additional information at test time (optical flow).

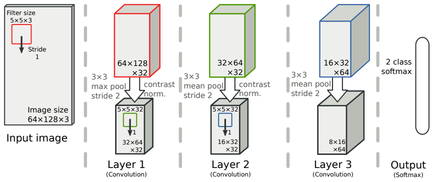

Figure 1: Illustration of the CifarNet, $\sim 10^5$ parameters.

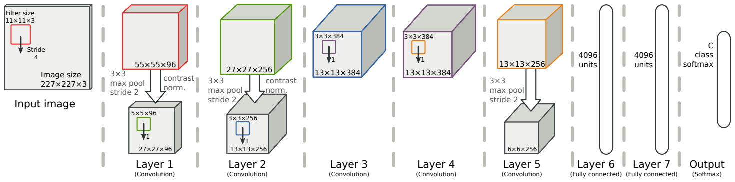

Figure 2: Illustration of the AlexNet architecture, $\sim 6 \cdot 10^7$ parameters.

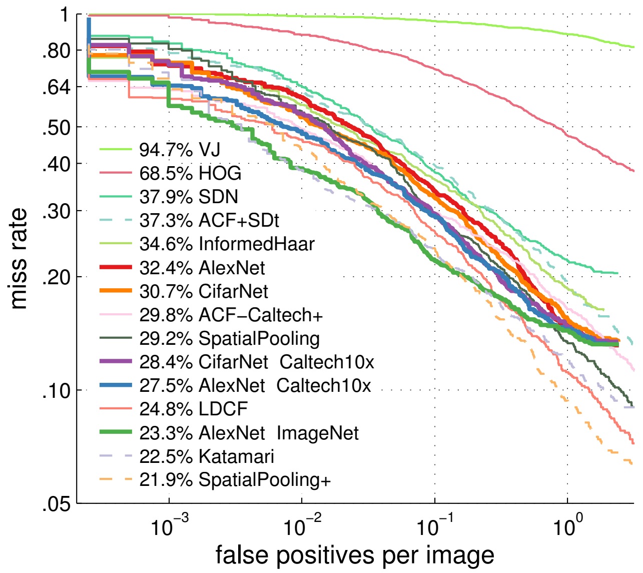

Figure 3: Comparison of our key results (thick lines) with published methods on Caltech test set. Methods using optical flow are dashed. Here lower is better.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014.

[3] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012.

[4] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR*, 2013.

[5] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[6] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, 2013.

[7] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 2013.