Pedestrian Detection using Deep Learning

R Santhosh Kumar
EE12B101

December 2015

Introduction

Pedestrian detection is the task of detecting humans with specific poses. This problem has been tackled using various techniques. The general principle is to extract features, and then train a classifier with these features as input. Deep learning architectures currently provide the top results for general object classification[2, 3], general object detection[4], scene recognition[5], pose estimation[6, 7] and many other problems. In this project, the problem of pedestrian detection has been tackled using deep learning.

Training data

CALTECH dataset

The dataset considered for this project is the Caltech Pedestrian Detection dataset[8, 9]. It consists of videos captured from a car traversing the U.S. streets under good weather conditions. The standard training set in the "Reasonable" setting consists of 4250 frames with $\sim 2.10^3$ annotated pedestrians, and the test set covers 4024 frames with $\sim 1.10^3$ pedestrians. There are 11 video sequences given. The first 5 videos are used for training, the 6^{th} video is used for validation and the remaining 5 videos are used for testing. Images are sampled from every 30^{th} frame.

CALTECH-10x dataset

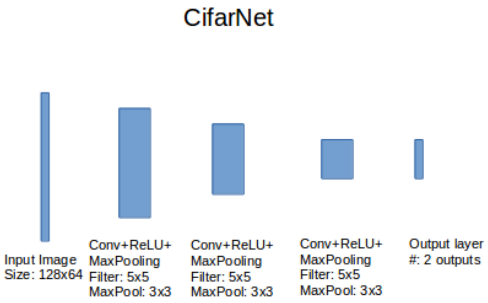
The Caltech Pedestrian Detection dataset is sampled at every 3^{rd} frame for the first 5 videos. The positives samples are extracted from every 3^{rd} frame, and the negative samples are extracted from every 30^{th} frame.

Deep Learning architectures

Three architectures have been demonstrated.

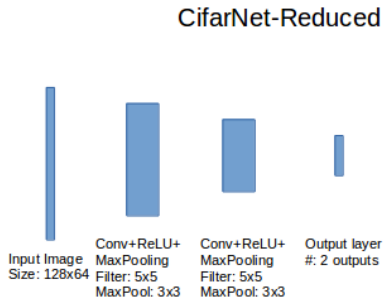
CifarNet (CN)

The CifarNet is a small 3 layer(convolutional) network which was designed to solve the CIFAR-10 classification problem.



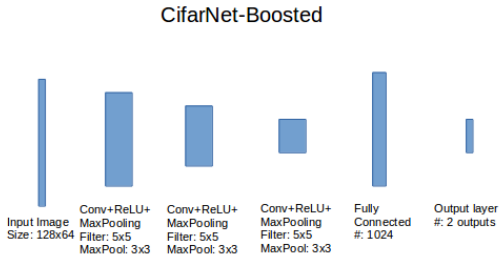
CifarNet-Reduced (CN-R)

The CifarNet-Reduced is a small 2 layer(convolutional) network derived from CifarNet. The final convolution layer is removed to obtain this network.



CifarNet-Boosted (CN-B)

The CifarNet-Boosted is a small 3 layer(convolutional) network derived from CifarNet. An extra fully connected layer is added before the classification layer.



Training pipeline

The general pipeline used for training the networks is explained here. Object proposals are extracted using SquaresChnFtrs[9, 10], an integral channel features based cascade classifier. It is used as a class-specific proposal generator, which reduces the search space.

Thresholds for positive and negative samples

The training proposals and ground truth annotations are taken. A proposal is said to be positive if it exceeds a certain Intersection over Union (IoU) threshold for atleast one GT annotation. It is considered negative if it does not exceed a second IoU threshold for any GT annotation, and is ignored otherwise. Both the thresholds were selected as 0.5. Also, ground truth annotations are added as positive samples.

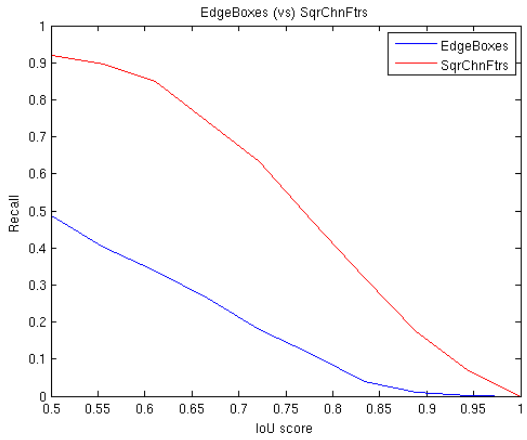
Training the network

After the positive and negative samples are obtained, the training data is created by ensuring that the ratio of positive and negative samples is maintained as a constant. Training and validation data are created. The validation data has equal number of positive and negative samples. Each of the architectures are trained from scratch, with random initialization to the weights. Negative Log Likelihood loss function is used for optimization. To deal with the data imbalance, the loss function for the negative samples is scaled down by the ratio of negative samples to positive samples.

Experiments

Proposal generator evaluation

SquaresChnFtrs is compared with EdgeBoxes[11]. The recall is computed for the two methods at different IoU values.

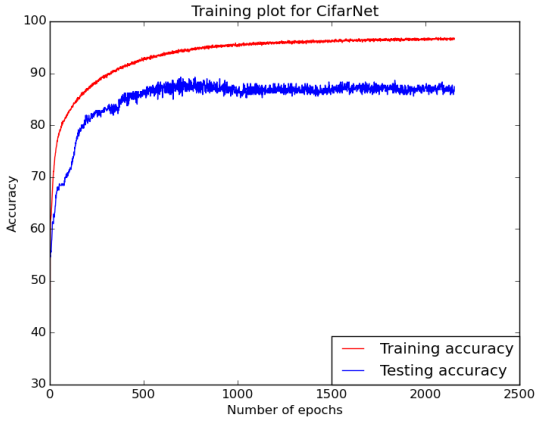


EdgeBoxes is a state-of-the-art category independent proposals generator. It can be clearly seen that SquaresChnFtrs is a better proposals generator for pedestrian detection.

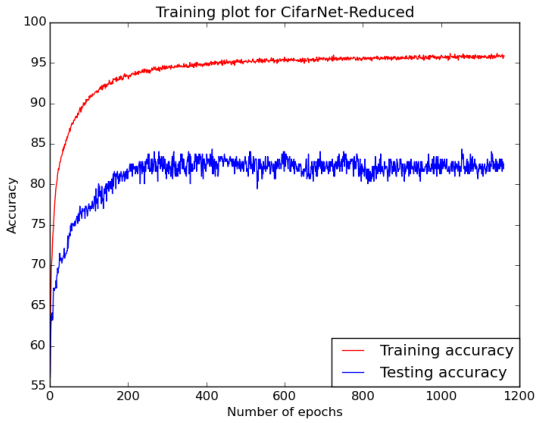
Network training

The training and validation accuracies are shown below for the different architectures.

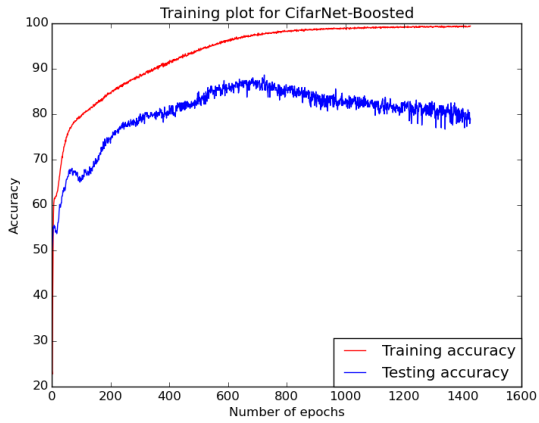
CifarNet



CifarNet-Reduced

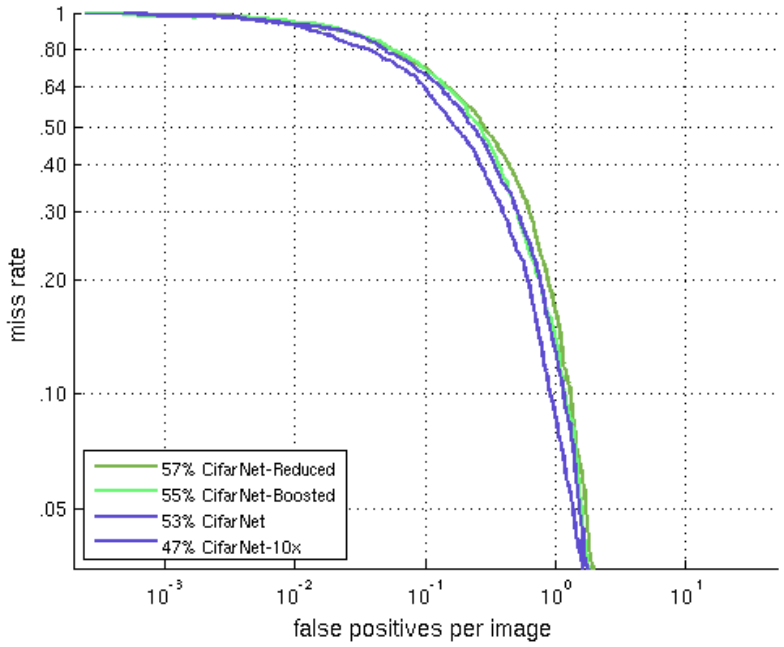


CifarNet-Boosted



Evaluation of trained networks

The trained networks are evaluated and the results are shown below. Miss rate is defined as $1 - recall$. Miss rate is plotted as a function of the False Positives Per Image. CifarNet-10x is the CifarNet trained on CALTECH-10x.



Conclusions

Based on the above results, we can see that CifarNet performs the best. CifarNet outperforms CifarNet-Reduced, which is expected. This is because larger networks usually perform better. However, CifarNet also outperforms CifarNet-Boosted, which actually has more parameters. This could be because the amount of training data was not sufficient to tune the extra parameters effectively. As the data increases the performance of the network also increases. This can be clearly seen by the decreased miss rate of CifarNet-10x in comparison to CifarNet.

References

[1] Hosang, Jan and Mohamed Omran and Benenson, Rodrigo and Schiele, Bernt. Taking a Deeper Look at Pedestrians. In CVPR, 2015

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Im- agenet classification with deep convolutional neural networks. In NIPS, 2012.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. In arXiv, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Ra- binovich. Going deeper with convolutions. In arXiv, 2014

[5] B. Zhou, J. Xiao, A. Lapedriza, A. Torralba, and A. Oliva. Learning deep features for scene recogni- tion using places database. In NIPS, 2014

[6] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In NIPS, 2014

[7] X. Chen and A. Yuille. Articulated pose estimation with image-dependent preference on pairwise rela- tions. In NIPS, 2014

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Ped- estrian detection: An evaluation of the state of the art. TPAMI, 2011

[9] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In ECCV, CVRSUAD workshop, 2014

[10] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In CVPR, 2013

[11] C. Lawrence Zitnick and Piotr Dollár Edge Boxes: Locating Object Proposals from Edges. In ECCV, 2014