Bootstrapping Face Detection with Hard Negative Examples

Shaohua Wan Zhijun Chen Tao Zhang Bo Zhang Kong-kat Wong
Xiaomi Inc.
{wanshaohua, chenzhijun, tao.zhang, zhangbo, kkwong}@xiaomi.com

August 9, 2016

Abstract

Recently significant performance improvement in face detection was made possible by deeply trained convolutional networks. In this report, a novel approach for training state-of-the-art face detector is described. The key is to exploit the idea of hard negative mining and iteratively update the Faster R-CNN based face detector with the hard negatives harvested from a large set of background examples. We demonstrate that our face detector outperforms state-of-the-art detectors on the FDDB dataset, which is the de facto standard for evaluating face detection algorithms.

1 Introduction

Face detection is one of the most widely researched topics in computer vision, and has found successful applications in face verification [16, 18], face tracking [14, 9], face recognition [11, 12], etc. However, it remains a challenging problem due to the appearance variations caused by changes in viewpoints, occlusion, facial expression, illumination and cosmetics, etc.

Since the remarkable success of the deep Convolutional Neural Network (CNN) [10] in image classification on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012, numerous efforts have been made to port CNN to face detection. Faster R-CNN [13], originally developed as a generic object detector, was recently demonstrated to produce state-of-the-art performance on face detection tasks [8].

Faster R-CNN is trained through a reduction that converts object detection into an image classification problem. That is, Faster R-CNN learns to detect objects by scoring a large set of regions of interest (RoIs) that are sampled from the images. This inevitably introduces a challenge that is not seen in standard image classification problems: the set of RoIs is distinguished by a highly imbalanced distribution; the ratio between the number of background (bg) RoIs (regions not belonging to any object of interest) and the number of foreground (fg) RoIs could reach as high as 100:1. Given such a high bg/fg ratio, it is virtually impossible to consider all the region proposals simultaneously. Faster R-CNN counteracts data imbalance by randomly subsampling the background examples and maintaining a bg/fg ratio of 3:1 in each mini-batch. This naive technique mitigates data imbalance to some extent, but more advanced strategy is required to further improve object detection performance.

Data imbalance has long been a problem for region-based object detectors. It is common that training data contains an overwhelming number of negative examples, most of which are easy ones for the detector and only a few are difficult. A standard solution, known as hard negative mining, is to iteratively grow, or bootstrap, a small set of negative examples by selecting those negatives for which the detector triggers a false positive alarm [17, 1, 2, 5]. This strategy leads to an iterative training algorithm that alternates between learning the detection model given the current set of examples, and then using the learned model to find hard negatives to add to the training set.

In this report, we adopt this simple yet effective hard negative mining technique for training state-of-the-art face detection models based on Faster R-CNN. We demonstrate that it yields significant boosts in detection performance on face detection benchmarks like FDDB [7].

2 Related Work

Hard negative mining was first introduced by Sung and Poggio [17] to select high quality examples for function approximation learning tasks. Since then, hard negative mining has been widely used to train region-based object detectors [1, 2, 4]. Recently significant performance boosts in object detection were made possible by CNN. CNN-based object detectors like R-CNN [4] and SPPnet [5] employ SVMs trained with hard negative mining to detect objects. [15] propose an online hard example mining algorithm for Fast R-CNN, which uses a data sampling strategy that favors those with high training loss. [19] adopt cascaded Boosted Forest, which perform effective hard negative mining and sample reweighting, to classify the region proposals generated by RPN. Our work is unique from previous works in that we harvest hard negative examples from the output of Fast R-CNN, which are used to jointly re-train both RPN and Fast R-CNN.

3 Bootstrapping Faster R-CNN with Hard Negative Examples

3.1 Overview of Faster R-CNN

Faster R-CNN consists of two modules. The first, called the Region Proposal Network (RPN), is a fully convolutional network for generating regions of interest (RoIs) that denote the possible presence of objects. The second module is the Fast R-CNN, whose purpose is to classify the RoIs and refine their position and scales. To save computing resources, RPN and Fast R-CNN share the same convolution layers up to their own fully connected layers. The mini-batch in each SGD step consists of foreground RoIs and background RoIs sampled from each image.

Foreground RoIs Foreground RoIs are those whose intersection over union (IoU) overlap with a ground-truth bounding box is greater than $\mathsf{th}_{\mathsf{fg}}$. $\mathsf{th}_{\mathsf{fg}} = 0.5$ is a fairly standard design choice, which finds widespread use in R-CNN, SPPnet, and Fast R-CNN. The same setting is used in Faster R-CNN.

Background RoIs Background RoIs are those whose maximum IoU with ground truth is in the interval $[\mathsf{th}_{\mathsf{bg}}, 0.5)$. $\mathsf{th}_{\mathsf{bg}} = 0.1$ is used by both Fast R-CNN and SPPnet. The idea is that regions with some overlap with the ground truth are more likely to be hard negatives [3].

Balancing bg-fg RoIs To handle the data imbalance, Faster R-CNN [3, 13] fix the bg/fg ratio in each mini-batch to a target of 3:1 by undersampling the bg RoIs at random. As mentioned in Section 1, detectors trained in this manner are susceptible to hard negative examples and tend to produce more false alarms.

3.2 Hard Negative Mining

The key idea of hard negative mining is to construct an initial training set consisting of positive RoIs and random subset of negative RoIs. The face detection model is learned with this training set and subsequently applied to all negatives to harvest false positives. The false positives are then added to the training set and then the model is trained again. This process is iterated several times until satisfaction.

We label a detected region as a hard negative if its maximum IoU with any groundtruth face annotation is less than 0.5. When the model is re-trained with the hard negatives added, we still maintain a bg/fg ratio of 3:1 in each mini-batch while ensuring that the hard negatives harvested by the previously trained model are selected.

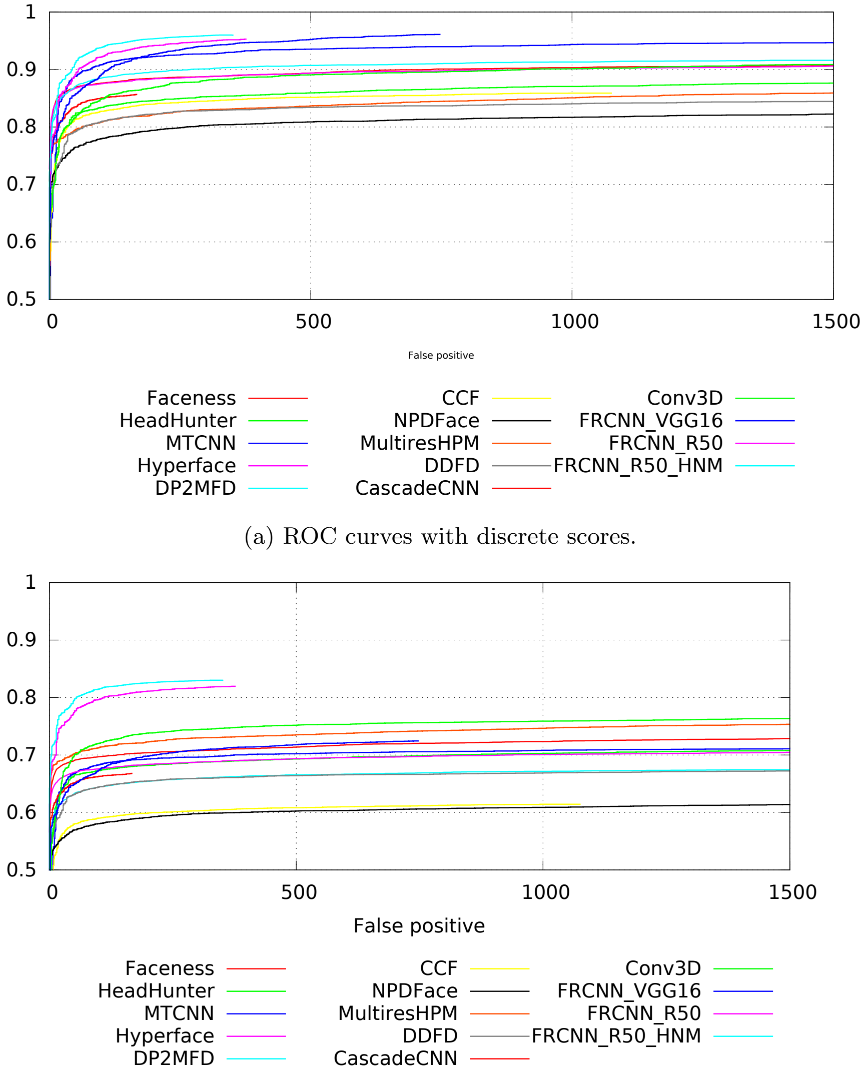
One should note that two optimization schemes exist for solving a Faster R-CNN network: (1) a 4-step alternating optimizaiton algorithm that alternates between fine-tuning for RPN and then fine-tuning Fast R-CNN; (2) an approximate joint optimization algorithm that solves Faster R-CNN as a single network. We take the latter optimization scheme for its simplicity and speed.

4 Experiments

In this section, we compare various detection algorithms and report their performance on the FDDB [7] face detection dataset.

4.1 FDDB

- FDDB has a total of 5,171 faces in 2,845 images, with a wide range of detection difficulties including occlusions, difficult poses, and low resolution and out-of-focus faces. As specified in [7], we conduct 10-fold cross-



(a) ROC curves with discrete scores.

(b) ROC curves with continuous scores.

Figure 1: The face detection performance of various methods on the FDDB dataset. FRCNN_VGG16 is the Faster R-CNN face detector with the VGG16 network architecture, as described in [8]. FRCNN_R50_HNM is the Faster R-CNN face detector with the ResNet-50 network architecture, trained with hard negative mining. We also obtain results for FRCNN_R50, which is the Faster R-CNN face detector with the ResNet-50 network architecture, trained without hard negative mining.

validation experiments, and use the FDDB evaluation software to generate two performance curves: (a) discrete ROC curve, and (b) continuous ROC curve. To generate the discrete ROC curve, each detection is assigned a binary match/non-match label. The continuous ROC curve associates a real-valued score with each detection based on the overlap between the detected and the annotated regions.

We train and test Faster R-CNN on single-scale images. The images are rescaled such that the shorter side is $s = 600$ pixels. ResNet-50 [6] is used as the network architecture and is initialized by the pre-trained ImageNet classification model. Approximate joint optimization mode is used to train Faster R-CNN, with a base learning rate of 0.001 for the first 50k iterations and a reduced learning rate of 0.0001 for 20k more iterations.

Our hard negative mining proceeds in two rounds. In the first round, we train the Faster R-CNN face detection model just as normal. In the second round, hard negatives are harvested from the training images using face detection model obtained from the first round. The face detection model is re-trained with the hard negatives added to each mini-batch.

We compare the proposed face detection algorithm with several other state-of-the-art algorithms listed on the official FDDB website, and plot the discrete and continuous ROC curves in Fig. 1. FRCNN_VGG16 is the Faster R-CNN face detector with the VGG16 network architecture, as described in [8]. FRCNN_R50_HNM is the Faster R-CNN face detector with the ResNet-50 network architecture, trained with hard negative mining. We also obtain results for FRCNN_R50, which is the Faster R-CNN face detector with the ResNet-50 network architecture, trained without hard negative mining. As can be seen, for both discrete and continuous ROC curves, FRCNN_VGG16, FRCNN_R50, and FRCNN_R50_HNM are the three best performing methods for face detection on FDDB, demonstrating the effectiveness of Faster R-CNN for face detection tasks. FRCNN_R50 outperforms FRCNN_VGG16 mainly due to the superiority of ResNet-50 over VGG16 in learning a deeper, better image feature representation. Moreover, hard negative mining is able to boost the performance of FRCNN_R50 by a large margin, as indicated by the ROC curve of FRCNN_R50_HNM. For detailed performance statistics, we refer readers to the official FDDB website [1].

5 Conclusion

In this report, we present the latest face detection results on the FDDB dataset, demonstrating that significant performance gains over state-of-the-arts can be achieved by learning Faster R-CNN face detection models with hard negative mining.

It is worth noting that, our face detection model, though being robust to difficult imaging conditions such as occlusions, difficult poses, and low-resolution

[1] http://vis-www.cs.umass.edu/fddb/results.html

faces, is less resilient to faces of small sizes. This is mainly due to the low-resolution feature maps of the last convolution layer on small faces. One possible solution is to pool feature maps of shallow but high-resolution convolution layers, and this is subject to future investigation.

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[2] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, September 2010.

[3] Ross Girshick. Fast R-CNN. In *International Conference on Computer Vision*, 2015.

[4] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, pages 346–361. 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *arXiv prepring arXiv:1506.01497*, 2015.

[7] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[8] Huaizu Jiang and Erik G. Learned-Miller. Face detection with the faster R-CNN. *CoRR*, abs/1606.03473, 2016.

[9] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry A. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012.

[11] K. Susheel Kumar, Vijay Bhaskar Semwal, and R. C. Tripathi. Real time face recognition using adaboost improved fast pca algorithm. *CoRR*, abs/1108.1353, 2011.

[12] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, 2015.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 91–99. 2015.

[14] Chi-Young Seong, Byung-Du Kang, Jong-Ho Kim, and Sang-Kyun Kim. Effective detector and kalman filter based robust face tracking system. In *Proceedings of the First Pacific Rim Conference on Advances in Image and Video Technology*, pages 453–462, 2006.

[15] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *CoRR*, abs/1604.03540, 2016.

[16] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[17] K.-K. Sung and T. Poggio. Learning and example se-lection for object and pattern detection. *MIT A.I. Memo*, 1521, 1994.

[18] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[19] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *arXiv prepring arXiv:1607.07032*, 2016.