A New Facial Detection Model based on the Faster R-CNN

Long Hao, F. Jiang

College of Electronic Engineering, Guangxi Normal University, Guilin 541004, China

fjiang@gxnu.edu.cn

Abstract. The object detection approaches in conjunction with Fast/Faster R-CNN and YOLO have shown the benchmarking performance on several occasions. Inspired by the Refine Net, we propose a new model called Faster+ R-CNN based on Faster R-CNN, which is mainly based on iterative refinement on the proposed regions. The Faster+ R-CNN model can iteratively refine the region proposal based on previous output. We trained and tested our new model on PASCAL VOC 2007 dataset, and experiments showed that our method can iteratively improve the mean average precision (mAP) from 0.6702 to 0.6764 in object detecting task. We also demonstrate the facial detection results using the Faster+ R-CNN on the widely used Face Detection Dataset and Benchmark (FDDB) benchmark. By training the Faster+ R-CNN model on the large scale WIDER face dataset, we report the improved results on two widely used face detection benchmarks including FDDB.

1. Introduction

The object detection is a computer vision task that aims to detect instances of semantic objects of certain classes in digital images (and videos). For a given image, the object detection system detects what objects are included and where they locate. It plays an important role in face detection, self-driving cars, video surveillance and many other applications. Object classification [13] and object detection [10] have rapidly progressed with advancements in convolutional neural networks (CNNs) [14] and the advent of large visual recognition datasets. Object detectors such as region proposal based methods like Fast/Faster R-CNN and regression based methods like YOLO [10] have shown the state-of- the art performance on several benchmarks.

The previous work is introduced in Section II, and then a new model called Faster+ R-CNN based on Faster R-CNN is introduced in Section III. In order to test the accuracy of the Faster+ R-CNN model, experiments show that our method can iteratively improve the mean average precision on PASCAL VOC 2007 dataset. Meanwhile, in order to prove its advantages in face recognition application, we report the experimental results on two widely used face detection bench-markings approach FDDB in Section IV.

2. Related Work

The classical object detection based on sliding windows and cascade methods achieves fast and reasonable accuracy on several applications such as face detection and human detection. They also demonstrated satisfactory results on high resolution images for detecting faces and license plates [3].

Girshick et al. [5] introduced a region-based CNN (R-CNN) for the object detection. The pipeline consists of two stages. In the first stage, a set of category-independent object proposals are generated, using selective search[14]. In the second refinement stage, the image region within each proposal is warped to a fixed size (e.g., 227 × 227 for the AlexNet [9]) and then mapped to a 4096-dimensional feature vector, which is fed into a classifier and also into a regressor that refines the position of the detection.

The Fast R-CNN [11] is a special case of the SPPnet, which uses a single spatial pyramid pooling layer, i.e., the region of interest (ROI) pooling layer, and thus allows end-to-end fine-tuning of a pre-trained ImageNet model. This is the key to its better performance relative to the original R-CNN. The most representative region proposal based mode is Faster R-CNN [11]，which originates from R-CNN [4].The Faster R-CNN demonstrates impressive results on various object detection benchmarks.

The key differences for the R-CNN, the Fast R-CNN, and the Faster R-CNN are summarized in Table 1. The running time of different modules are reported on the FDDB dataset [7], where the typical resolution of an image is about 350 × 450. We can clearly see that the entire running time of the Faster R-CNN is significantly lower than that for both the R-CNN and the Fast R-CNN.

		R-CNN	Fast R-CNN	Faster
proposal stage	time	Faceness[15]:9.91s(+2.73s=12.64s)		0.32s
		EdgeBox:2.73s		
		DeepBox[10]:0.27s(+2.73s=3.00s)		
refinement	input to	cropped proposal	input image &	input
	#forward	#proposals	1	1
	time	7.04s	0.21s	0.06s
T Total	time	R-CNN + Edge Box: 9.77s	Fast R-CNN + EdgeBox: 2.94s	0.38s
		R-CNN + Faceness: 19.68s	Fast R-CNN + Faceness: 12.85s	
		CNN + DeepBox: 10.04s	R-CNN + DeepBox: 3.21s	

Figure 1. Comparisons of the entire pipeline of different region-based object detection methods.

3. Refining Faster R-CNN (Faster + R-CNN)

3.1. Faster R-CNN

One of the state-of-the-art object detection models is called Faster R-CNN [11]. The architecture of Faster R-CNN is shown in Figure 2, which is composed of 5 main parts: a deep fully convolutional network; region proposal network; ROI pooling and fully connected networks; bounding box regressor; and the classifier. For a given image, we first employ a pre-trained deep convolutional neural network, such as VGG [13], to extract feature maps from it. Then, they use a Region Proposal Network (RPN), which consists of two convolutional layers, to detect the regions that might contain object in the feature maps (image). Then the network employs a ROI pooling layer [3] to crop and resizes the feature maps according to these region proposals. The new feature maps of each region are then used for classification and finer bounding box regression through three fully connected layers.

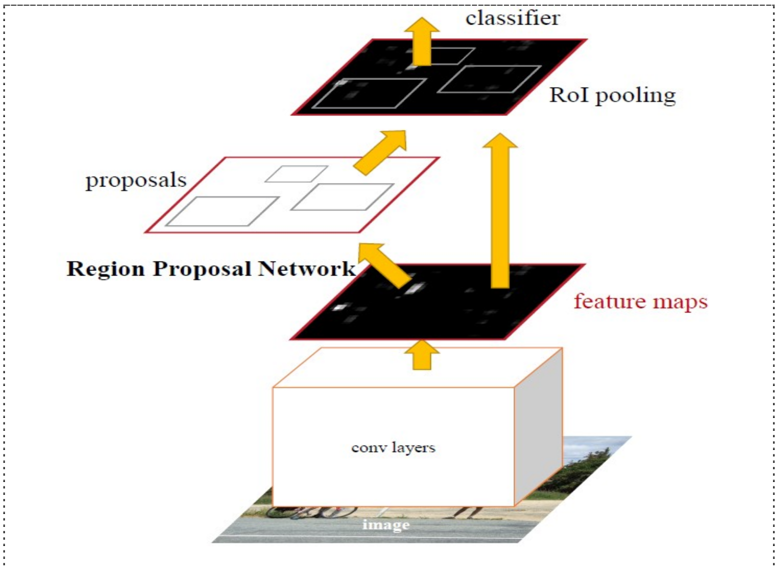

Figure 2. Architecture of Faster R-CNN. (This figure is from the original paper [11])

3.2. *Faster + R-CNN*

Inspired by the Refine Net [3], we propose a new model called Faster+ R-CNN based on Faster R-CNN, which is mainly based on iterative refinement on the proposed regions. The model of Faster+ R-CNN is shown in Figure 2. In the first iteration, it's exactly the same as Faster R-CNN: extract feature maps from image by VGG, pass them to RPN to get region proposals, employ ROI pooling layer to crop and resize new feature maps for each proposal, and use a three-layer fully connected network to get the final class scores and bounding box regression for each class.

In the second iteration, we select the regressed bounding box with the maximum class score as the region proposal in the second iteration. And the rest of second iteration is the same as the first iteration: we use ROI pooling layer to crop and resize feature maps for each proposal, classify and regress the bounding box using new feature maps. Also, we reuse parameters of the three-layer fully connected network in different iterations. After the second iteration, we can repeat the same process for the third iteration, and so on. Figure 3 is a demonstration of iteration number and iteration number is 3.

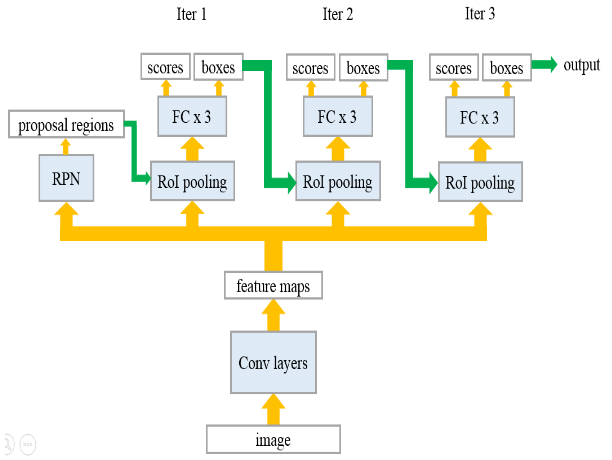

Figure 3. Architecture of Faster R-CNN with iterative region proposal refinement.

3.3. The VOC 2007 data validation

We implemented our models based on a PyTorch implementation of Faster+ R-CNN and trained our iterative refinement model on VOC 2007 training and validation datasets, and tested our models on VOC 2007 test set. We use mean average precision (mAP), which is widely used for object detection task, as the metric to evaluate our model. To better understand our iterative refinement model and to analyze its pros and cons, we visualize some object detection results in Fig 4, where refinement with iteration 1, 2 and 3 are shown in bounding box with color orange, yellow and green respectively. We also compare the performance of the Fast R-CNN, the Faster R-CNN and the Faster+ R-CNN on the VOC 2007 data from Table 2 and Table 3 respectively. As can be observed from Fig. 3, the Faster+ R-CNN significantly outperforms the other two methods.

model	**parameters**				**mAP**
	train max	**test max**	**includes RPN loss**	**learning rate**	
Faster R-CNN	-	-	-	-	0.6702
Faster+ R-CNN	2	2	Y	0.0001	0.6580
Faster+ R-CNN	2	2	Y	0.00001	0.6707
Faster+ R-CNN	3	3	N	0.00001	0.6744
Faster+ R-CNN	3	2	Y	0.00001	0.6762
Faster+ R-CNN	3	3	Y	0.00001	**0.6764**
Faster+ R-CNN	3	4	Y	0.00001	0.6760

Figure 3. Architecture of Faster R-CNN with iterative region proposal refinement.

From Figure 3 we can observe iterative improvement of the bounding box predictions in some images. More concretely:

1) In images with single object where the original region proposal only covers a part of the object: the refinement can usually enlarge the bounding box to cover the entire object, as shown in 2A, 3A, 3B and 5A in Figure 3.

2) In images with single object where the original region proposal is already very good: the refinement usually makes minor changes to the bounding box, as shown in 2C.

3) In images with multiple object where region proposal for some objects are too large, iterative refinement can sometimes shrink it to the correct size (as in 1A, 1B and 4A) and sometimes fail to do so.

4) Most importantly, the iterative refinement process can some- times find object that was not originally detected, as shown in 1C (where the truck under the plane is not detected until the third refinement iteration) and 3C (where the dog is not detected until the second iteration).

4. Comparison and verification of different methods

4.1. The WIDER face dataset

We train a Faster+ R-CNN face detection model on WIDER face dataset [16]. There are 12,880 images and 159,424 faces in the training set. We train the face detection model based on a pre-trained Image Net model, VGG16 [13]. We randomly sample one image per batch for training. In order to fit it in the GPU memory, it is resized based on the ratio 1024/ max (w, h), where w and h are the width and height of the image, respectively. We run the SGD solver 50k iterations with a base learning rate of 0.001 and run another 20K iterations reducing the base learning rate to 0.00012

We test the trained face detection model on datasets FDDB [7] . There are 10 splits in FDDB. For testing, we resize the input image based on the ratio min (600/ min(w, h), 1024/ max(w, h)). For the RPN, we use only the top 300 face proposals to balance efficiency and accuracy.

4.2. Comparison of different methods

We also compare face detection performance of the Fast R-CNN, the Faster R-CNN and the Faster+ R-CNN on FDDB. For the R-CNN, Fast R-CNN and the Faster+ R-CNN, we use the top 2000 proposals generated by the Faceness method [15]. As can be observed from Fig. 3, the Faster+ R-CNN significantly outperforms the other two methods. In Fig. 4, we demonstrate some randomly sampled images of the WIDER. The sample detection results are on the FDDB datasets where green are ground-truth annotations and red are detection results of the Faster+ R-CNN.

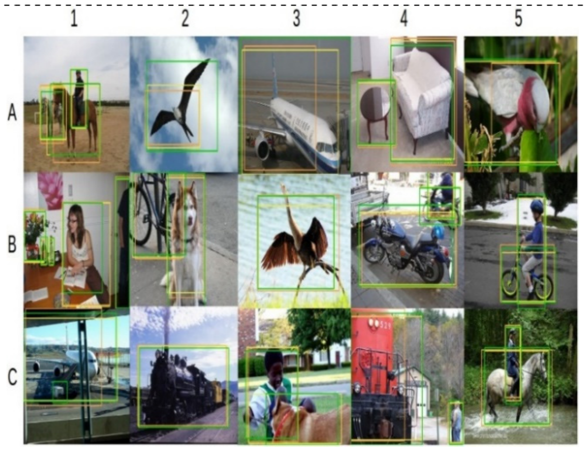

Figure 4. Visualization of iterative refinement model (Iteration 1, 2, and 3 are denoted by orange, yellow and green)

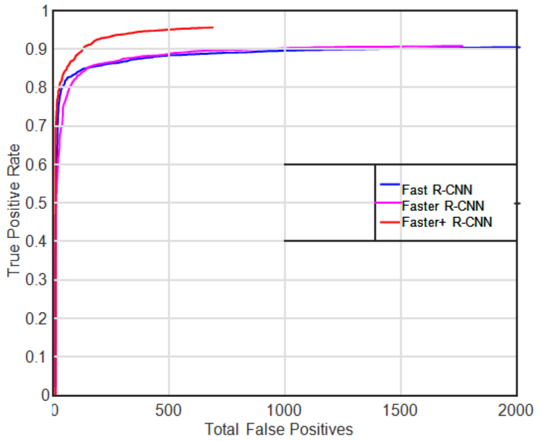

Figure 5. Comparisons of region-based CNN object detection methods for face detection on FDDB.

5. Conclusion

In this paper, we demonstrate the face detection results using the Faster+ R-CNN on Face Detection Dataset and Benchmark (FDDB). We also compare different generations of region-based CNN object detection models, and compare to a variety of other recent high-performing detectors. Experimental results suggest that its effectiveness comes from the region proposal network (RPN) module. Due to the sharing of convolutional layers between the RPN and Fast R-CNN detector module, it is possible to use a deep CNN in RPN without extra computational burden. Although the Faster+ R-CNN is designed for generic object detection, it demonstrates impressive face detection performance when retrained on a suitable face detection training set. It may be possible to further boost its performance by considering the special patterns of human faces.

Acknowledgments

Firstly the authors appreciate the help from the anonymous reviewers on their valuable advice and suggestions in improving the final quality of this paper. Secondly, the authors would like to thank the colleagues and classmates in the laboratory, for the timely help/assistances and encouragements all the time. This work is partially supported by the Guangxi Program (F-KA16016), National Natural Science Foundation of China (61762018) and the Colleges and Universities Key Laboratory of Intelligent Integrated Automation, Guilin University of Electronic Technology, China (GXZDSY2016-03), the research fund of Guangxi Key Lab of Multi-source Information Mining & Security (18-A-02-02).

References

[1] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recog- nition with visual attention. CoRR, abs/1412.7755, 2014.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303– 338, 2010.

[3] R. Girshick. Fast r-cnn. In Proceedings of the IEEE Inter- national Conference on Computer Vision, pages 1440–1448, 2015.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fea- ture hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[5] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In European Conference on Computer Vision, pages 21–37. Springer, 2016.

[8] Longcw. Faster rcnn with pytorch. https://github.com/longcw/faster_rcnn_pytorch, 2017.

[9] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi. Refinenet: Iterative refinement for accurate object localization. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pages 1528–1533, Nov 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 779–788, 2016.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.

[13] K. Simonyan and A. Zisserman. Very deep convolu- tional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[14] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. Interna- tional journal of computer vision, 104(2):154–171, 2013.

[15] C. Xiong, V. Zhong, and R. Socher. Dynamic coattention networks for question answering. CoRR, abs/1611.01604,2016.