Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Summary written by Mahban Gholijafari
March 25, 2025

Summary: This paper introduces Region Proposal Network (RPN) that shares full-image convolution features with the detection network and enables nearly cost-free region proposals. RPNs can generate high-quality region proposals and are used by Fast R-CNN for detection.

Related work: This work is computing proposals with a deep net and several recent papers have proposed ways of using deep networks for locating class-specific or class-agnostic bounding boxes [8, 5, 2, 7]. For efficient and accurate visual recognition, shared computation of convolutions has been proposed [5, 4, 1, 3].

Approach: A RPN takes an image of any size as input and outputs a set of rectangular object proposals that have an objectness score. To do so, this work slides a small network that is fully connecten to a $n \times n$ spatial window of the input conv feature map over the conv feature map output by the shared conv layer between Fast R-CNN and RPN. Each sliding window is mapped to a lower-dimensional vector that is fed into two sibling fully-connected layers, a box-regression layer (*reg*) and a box-classification layer (*cls*). k region proposals are simultaneously predicted at each sliding-window location and therefore, the *reg* layer has $4k$ outputs encoding the coordinates of k boxes. The *cls* layer outputs $2k$ scores that estimate probability of object/not-object for each proposal and k proposals are parameterized relative to k reference boxes called anchors that are associated with a scale and aspect ratio. anchors and functions that compute proposals relative to anchors are translate invariant. A binary class label has been assigned to each anchor for training RPNs and this labels depends on Intersection-over-Union (IoU) overlap with a ground-truth box. The optimized loss function used in this paper is: $L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*)$, where i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground-truth label p_i^* is 1 if the anchor is positive, and is 0 if the anchor is negative. t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and t_i^* is that of the ground-truth box associated with a positive anchor. The classification loss L_{cls} os log loss over two classes and for regression loss, $L_{reg}(t_i, t_i^*) = R(t_i - t_i^*)$ is used, where R is the robust loss function. The outputs of the cls and reg layers are normalized with N_{cls} and N_{reg}, and a balancing weight λ. There are 4 steps in total for the proposed training algorithm. First, training the RPN as described, second, training a separate detection network by Fast R-CNN using the proposals generated by the step-1 RPN. Third, a detector network is used to initialize RPN training with fixed shared conv layers and finally, keeping the shared conv layers fixed, we fine-tune the fc layers of the Fast R-CNN and therefore, bot network will share the same conv layers and form a unified network.

Datasets, Experiments and Results: This work uses 3×3 spatial window followed by ReLUs applied to 171 and 228 pixels of Zeiler and Fergus model (ZF) [9] and Simonyan and Zisserman model (VGG) [6]. 3 scales and 3 aspect ratios have been used resulting in $k = 9$ anchors at each sliding position. To compute the loss function of a mini-batch, 256 anchors are randomly sampled in an image and the positive and the sampled negative anchors have a ratio of up to 1:1. All the new layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01 and other layers are initialized by pre-training model for ImageNet classification. Learning rate of 0.001 for 60k mini-batches, and 0.0001 for the next 20k mini-batches on the PASCAL dataset. The momentum of 0.9 and a weight decay of 0.0005 has been used and the implementation uses Caffe. Some RPN proposals highly overlap with each other and to reduce this redundancy, they proposed using non maximum suppression (NMS) based on their *cls* scores. The proposed method has been evaluated on PASCAL VOC 2007 detection benchmark that consists of 5k training images and 5k test images over 20 object categories.

Strengths: Unlike other works, to account for varying sizes of regression, k bounding-box regressors are learned, which will enable this technique to predict boxes of various sizes even though the features are of a fixed size/scale. The proposed method makes the region proposal step nearly cost-free and enables a unified deep learning-based object detection system to run at 5-17 fps. The learned RPN also improves the quality of the region proposal and this overall object detection accuracy. Compared to MultiBox, RPN uses fewer parameters and therefore reduces the risk of overfitting.

Weaknesses: The alternating optimization process between RPN and Fast R-CNN adds complexity to training. Crossing anchors at the image boundary are ignored to avoid large errors.

Reflections: The proposed method uses single-scale im-

ages for feature extraction and multi-scale feature extraction could improve proposal quality.

References

[1] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3992–4000, 2015.

[2] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[3] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[5] 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.

[6] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[7] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[8] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013.

[9] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.