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

SongSeungWu

1 ABSTRACT

Faster R-CNN addresses a key bottleneck in object detection—the region proposal step—by introducing the Region Proposal Network, which generates high-quality region proposals with almost no additional cost. RPN shares convolutional features with the detection network, predicting object bounds and objectness scores at each position and is trained end-to-end for optimal proposal generation. Integrated with Fast R-CNN, RPN functions as an "attention" mechanism, guiding the network on where to focus. Using the VGG-16 model, Faster R-CNN achieves state-of-the-art object detection at 5 frames per second on a GPU across datasets like PASCAL VOC and MS COCO, ranking first in multiple tracks at the ILSVRC and COCO 2015 competitions. The code is publicly available.

2 INTRODUCTION

Recent advances in object detection have largely relied on region proposal methods, such as Selective Search, combined with Region-based Convolutional Neural Networks. However, traditional region proposal methods are typically slower than detection networks, especially on the CPU, leading to significant bottlenecks. Faster R-CNN introduces RPN to overcome this issue by sharing convolutional layers with the detection network, enabling cost-free region proposals with minimal computation. RPN is a fully convolutional network that predicts object bounds and scores, creating proposals directly from convolutional feature maps at about 10ms per image on a GPU. RPN also incorporates anchor boxes as reference points for generating multi-scale region proposals, eliminating the need for costly image or filter pyramids. Faster R-CNN unifies RPN and Fast R-CNN into a single network, using alternating fine-tuning to rapidly converge. This integrated structure has shown strong results on PASCAL VOC and MS COCO, reaching 5 fps and ranking first in several ILSVRC and COCO 2015 categories. The model is widely applicable in research and industry for efficient, high-accuracy object detection.

3 FASTER R-CNN

Faster R-CNN consists of two primary modules: the Region Proposal Network and the Fast R-CNN detector. RPN generates region proposals, and Fast R-CNN detects objects using these proposals. Together, they form a unified detection network, where RPN guides Fast R-CNN on where to look through an attention mechanism. This structure allows shared features between the two modules for efficient training.

4 REGION PROPOSAL NETWORKS(RPN)

RPN generates candidate regions and assigns an objectness score to each, sharing convolutional layers with Fast R-CNN for efficiency. Anchors: RPN uses a sliding window approach, introducing anchors as reference boxes for different scales and aspect ratios. This ensures translation invariance and reduces the model's parameter count, lowering overfitting risks. Multi-Scale Anchors: Traditional methods rely on image or filter pyramids for multi-scale detection, while RPN uses multiple anchors on a single image and filter size to achieve efficiency. Loss Function: RPN uses a multi-task loss function, calculating classification and regression losses for each anchor based on object presence, enhancing proposal accuracy. Training: RPN is trained end-to-end using backpropagation and SGD. Positive and negative anchors are sampled from each image, with new layers initialized randomly and shared layers initialized from an ImageNet-pre-trained model.

5 SHARING FEATURES FOR RPN AND FAST R-CNN

- Alternating Training: Train RPN first, then use its proposals to train Fast R-CNN, iteratively refining both.
- Approximate Joint Training: Treat RPN proposals as fixed for Fast R-CNN training, combining shared layer signals for loss. This reduces training time by 25-50%.

6 EXPERIMENTS

- Experiments on PASCAL VOC: Faster R-CNN was evaluated on the PASCAL VOC 2007 and 2012 benchmarks. Using the ZF and VGG-16 models, RPN+Fast R-CNN demonstrated higher performance and speed compared to traditional methods like Selective Search and EdgeBoxes. Due to shared convolutional computations between RPN and Fast R-CNN, RPN+Fast R-CNN proved faster and more efficient than SS or EB, achieving a mean Average Precision of 69.9% with VGG-16.
- Ablation Experiments: The impact of RPN's cls and reg layers on detection accuracy was analyzed. The experiments showed that high mAP could be maintained with only the top 300 region proposals, and that using Non-Maximum Suppression to reduce redundant regions did not negatively impact detection accuracy. Adjusting anchor sizes and ratios significantly improved accuracy, and RPN's performance was further enhanced when paired with a more powerful VGG-16 model.
- Experiments on MS COCO: Faster R-CNN also demonstrated strong performance on the MS COCO dataset. With VGG-16, the model achieved 42.1% mAP@0.5 and 21.5% mAP@[.5, .95], and these metrics improved with ResNet-101. Faster R-CNN achieved top results in the COCO 2015 competition.
- Comparison of One-Stage vs. Two-Stage Detection Approaches: Faster R-CNN's two-stage approach consists of generating class-agnostic region proposals followed by object detection, achieving 4.8% higher mAP than the one-stage OverFeat approach. This experiment highlights that cascaded proposals lead to more accurate detections.
- Transfer Learning from COCO to PASCAL VOC: When the model trained on MS COCO was transferred to PASCAL VOC, it outperformed models trained solely on VOC data. The model pre-trained on COCO and evaluated directly on VOC achieved 76.1% mAP, higher than the 73.2% obtained with VOC07+12 training alone. Fine-tuning the COCO-trained model on VOC data further increased the mAP to 78.8%, demonstrating the significant boost that large-scale data and transfer learning can provide to object detection performance.

7 CONCLUSION

Faster R-CNN introduces Region Proposal Networks for efficient region proposal generation by sharing convolutional features with the detection network, resulting in minimal additional cost. This enables near real-time deep-learning-based object detection, improving both proposal quality and overall detection accuracy.