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

- By Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun

Presented by - Avirat Belekar

Overview

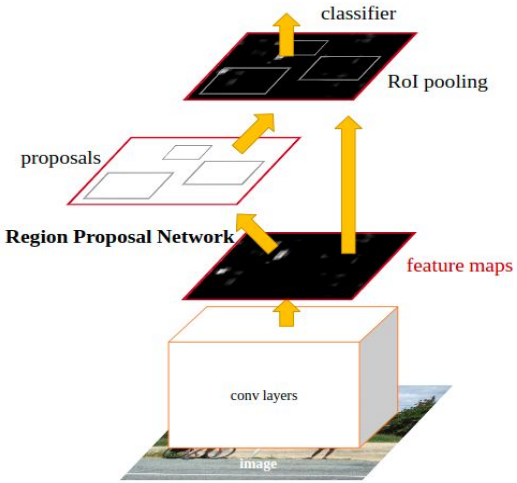
- Introduction
- Model Architecture
- Training Methodology
- Experimental Results
- Conclusions
- Questions

Introduction

- Object Detection is driven by the success of **Region Proposal Methods** and **Region based Convolutional Neural Networks**
- The computation of Proposals is a bottleneck in the state-of-art detection system
- In this paper, they have introduced Regional Proposal Network(RPN) that can share convolution layers with object detection networks, thus could reduce computational time

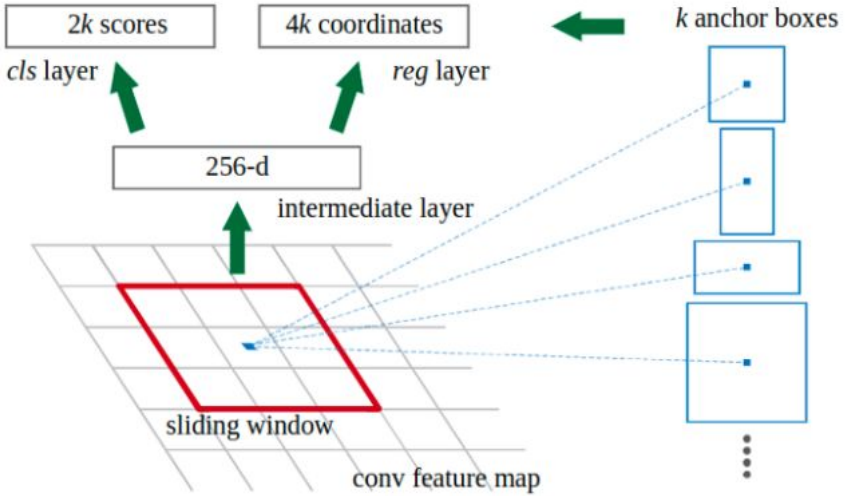
Model Architecture

1. Region Proposal Network
2. Fast R-CNN detector

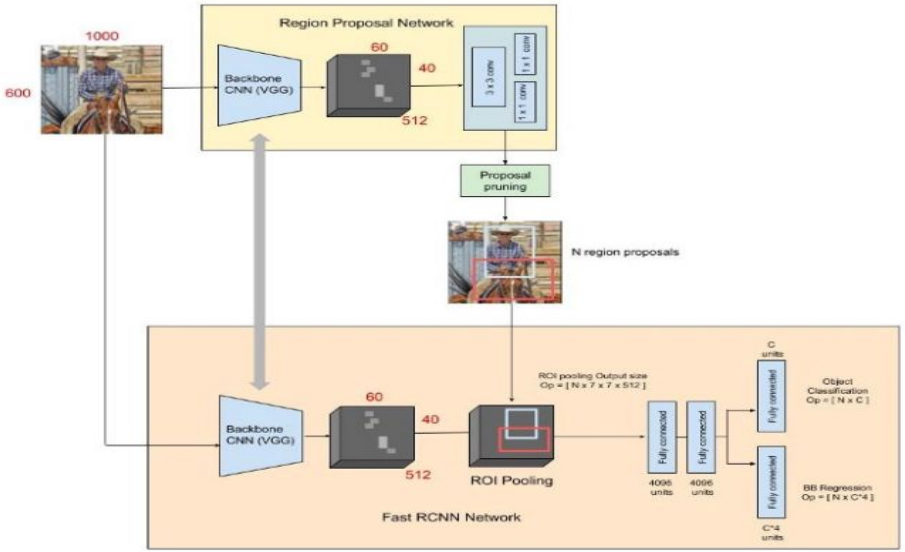


Model Architecture (2)

Anchor is another important technique used in this work. At each sliding window location, they simultaneously predict multiple region proposals where the number of maximum possible proposals for each location is denoted by k



Model Architecture (3)



Training Methodology

- The output feature map consists of about 40 x 60 locations, corresponding to 40*60*9 ~ 20k anchors in total
- An anchor is considered to be a “positive” sample if it satisfies either of the two conditions
  - The anchor has the highest IoU (Intersection over Union, a measure of overlap) with a ground truth box
  - The anchor has an IoU greater than 0.7 with any ground truth box
- An anchor is labeled “negative” if its IoU with all ground truth boxes is less than 0.3
- Each mini-batch for training the RPN comes from a single image.

Training Methodology Cont.

Loss Function :

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*).$$

Here i is the index of the anchor in the mini batch. The classification loss $L_{cls}(p_i, p_i^*)$ is the` log loss over two classes (object vs not object). p_i is the output score from the classification branch for anchor i, and p_i^* is the ground truth label (1 or 0).

The regression loss $L_{re}(t_i, t_i^*)$ is activated only if the anchor actually contains an object i.e., the ground truth p_i^* is 1. The term t_i is the output prediction of the regression layer and consists of 4 variables [t_x, t_y, tw, th].

Experimental Results

In their work , authors have used PASCAL VOC 2007 as their benchmark for detection. This database contains about 5k images for training over 20 object categories.

The image on the left depicts detection results on PASCAL VOC 2007 test results

train-time region proposals		test-time region proposals		
method	# boxes	method	# proposals	mAP (%)
SS	2000	SS	2000	58.7
EB	2000	EB	2000	58.6
RPN+ZF, shared	2000	RPN+ZF, shared	300	**59.9**
ablation experiments follow below				
RPN+ZF, unshared	2000	RPN+ZF, unshared	300	58.7
SS	2000	RPN+ZF	100	55.1
SS	2000	RPN+ZF	300	56.8
SS	2000	RPN+ZF	1000	56.3
SS	2000	RPN+ZF (no NMS)	6000	55.2
SS	2000	RPN+ZF (no *cls*)	100	44.6
SS	2000	RPN+ZF (no *cls*)	300	51.4
SS	2000	RPN+ZF (no *cls*)	1000	55.8
SS	2000	RPN+ZF (no *reg*)	300	52.1
SS	2000	RPN+ZF (no *reg*)	1000	51.3
SS	2000	RPN+VGG	300	59.2

Conclusions

This work presented by RPNs for efficient and accurate region proposal generation. By sharing features with the downstream detection network , the region proposal in nearly cost free

Future Scope: How well does Faster R-CNN work with multi scale images since the paper discusses only single scaled images?

Questions!

Thank You!