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

NeurIPS 2015

Shaoqing Ren[1], Kaiming He[2], Ross Girshick[3], Jian Sun[2]

[1]University of Science and Technology of China

[2]Visual Computing Group, Microsoft Research

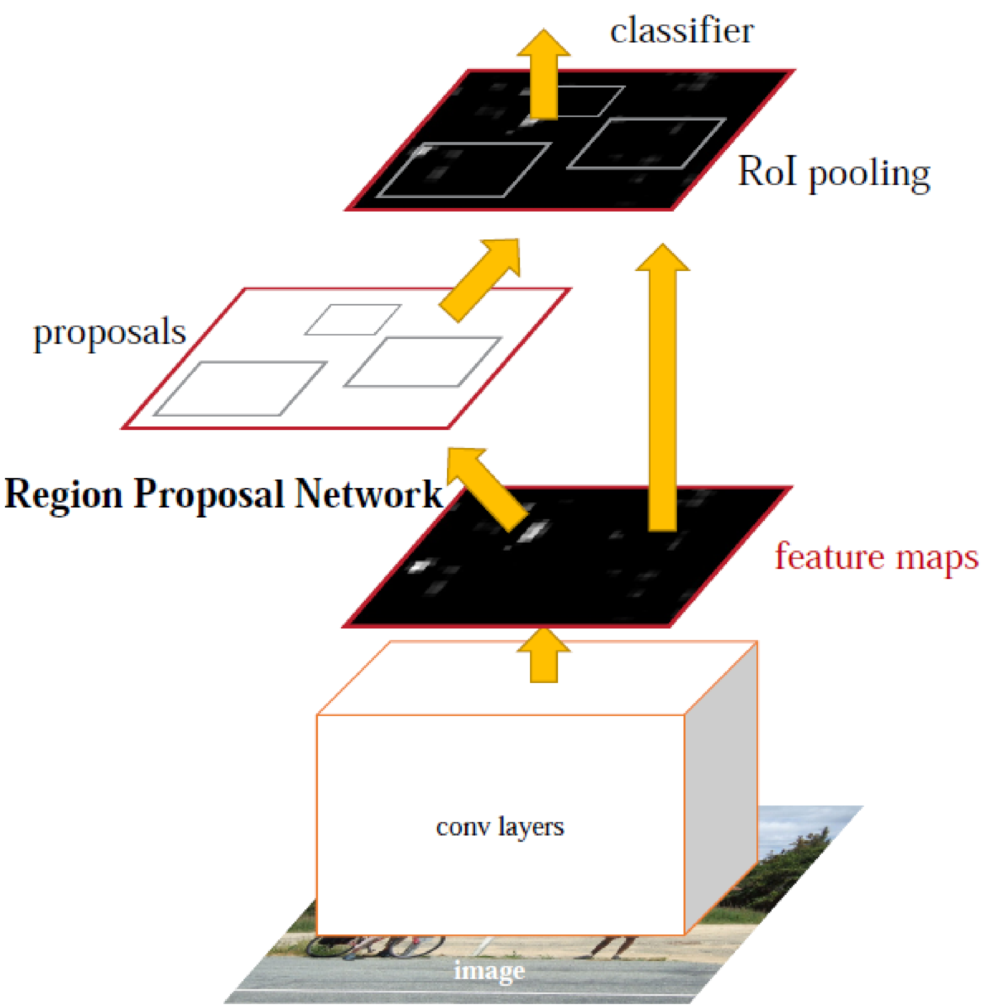
[3]Facebook AI Research

Introduction

- **Object Detection**
  - Recent advances
    - The success of region proposal methods (e.g., [4]) and region-based convolutional neural networks (RCNNs) [5]
  - Region-based CNNs
    - Computationally expensive as originally developed in [5]
    - Their cost has been drastically reduced thanks to sharing convolutions across proposals
  - Recent advances
    - The success of region proposal methods (e.g., [4]) and region-based convolutional neural networks (RCNNs) [5].

Introduction

A single, unified network for object detection



- **FASTER R-CNN**
  - The first module
    - A deep fully convolutional network that proposes regions
  - The second module
    - The Fast R-CNN detector that uses the proposed regions
  - The entire system is a single, unified network for object detection
  - Using the recently popular terminology of neural networks with 'attention' mechanisms, the RPN module tells the Fast R-CNN module where to look.

Method: A Region Proposal Network (RPN)

- **A Region Proposal Network (RPN)**
  - Takes an image (of any size) as input and outputs a set of rectangular object proposals, each with an objectness score
  - Model this process with a fully convolutional network
    - Share computation with a Fast R-CNN object detection network
    - Assume that both nets share a common set of convolutional layers
  - Generate region proposals
    - Slide a small network over the convolutional feature map output by the last shared convolutional layer
    - Takes as input an $n \times n$ spatial window of the input convolutional feature map.
  - Each sliding window is mapped to a lower-dimensional feature
    - This feature is fed into two sibling fullyconnected layers—a box-regression layer (reg) and a box-classification layer (cls)

Method: A Region Proposal Network (RPN)

- **Anchors**
  - At each sliding-window location, we simultaneously predict multiple region proposals
    - The number of maximum possible proposals for each location is denoted as k

  - The reg layer has $4k$ outputs encoding the coordinates of k boxes
  - The cls layer outputs $2k$ scores that estimate probability of object or not object for each proposal
  - The k proposals are parameterized relative to k reference boxes, which we call anchors

  - An anchor is centered at the sliding window in question
  - An anchor is associated with a scale and aspect ratio
  - Use 3 scales and 3 aspect ratios, yielding $k = 9$ anchors at each sliding position

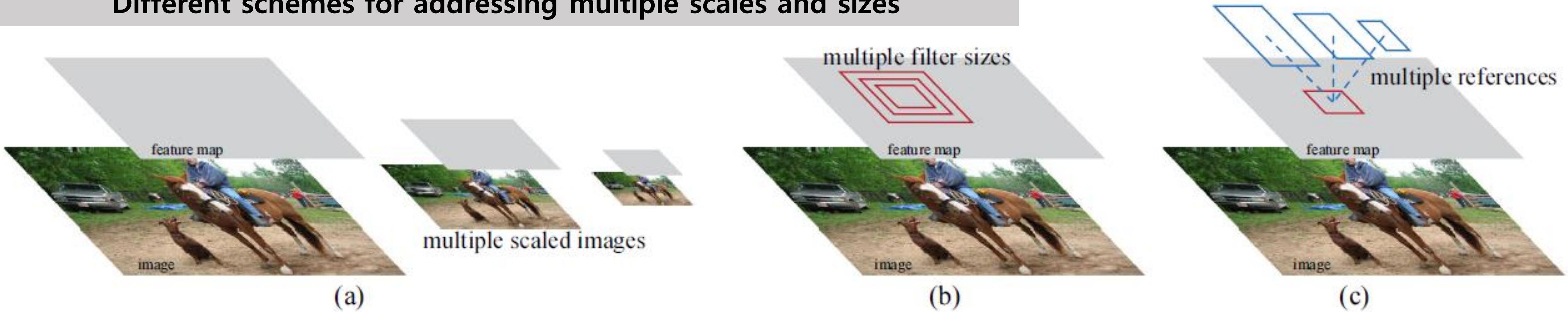
Method: A Region Proposal Network (RPN)

- **Translation-Invariant Anchors**
  - An important property of our approach is that it is translation invariant, both in terms of the anchors and the functions that compute proposals relative to the anchors
  - The translation-invariant property also reduces the model size

Method: A Region Proposal Network (RPN)

- **Multi-Scale Anchors as Regression References**
  - Our design of anchors presents a novel scheme for addressing multiple scales size
  - As a comparison, our anchor-based method is built on a pyramid of anchors, which is more cost-efficient
  - Our method classifies and regresses bounding boxes with reference to anchor boxes of multiple scales and aspect ratios

Different schemes for addressing multiple scales and sizes



- (a) Pyramids of images and feature maps are built, and the classifier is run at all scales
- (b) Pyramids of filters with multiple scales/sizes are run on the feature map
- (c) We use pyramids of reference boxes in the regression functions

Method

- **Loss Function**
  - For training RPNs, we assign a binary class label to each anchor
    - (i) The anchor/anchors with the highest Intersection-over- Union (IoU) overlap with a ground-truth box
    - (ii) An anchor that has an IoU overlap higher than 0.7 with any ground-truth box
  - A single ground-truth box may assign positive labels to multiple anchors
    - Adopt the first condition for the reason that in some rare cases the second condition may find no positive sample
  - Assign a negative label to a non-positive anchor if its IoU ratio is lower than 0.3 for all ground-truth boxes
    - Anchors that are neither positive nor negative do not contribute to the training objective

The multi-task loss in Fast R-CNN

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