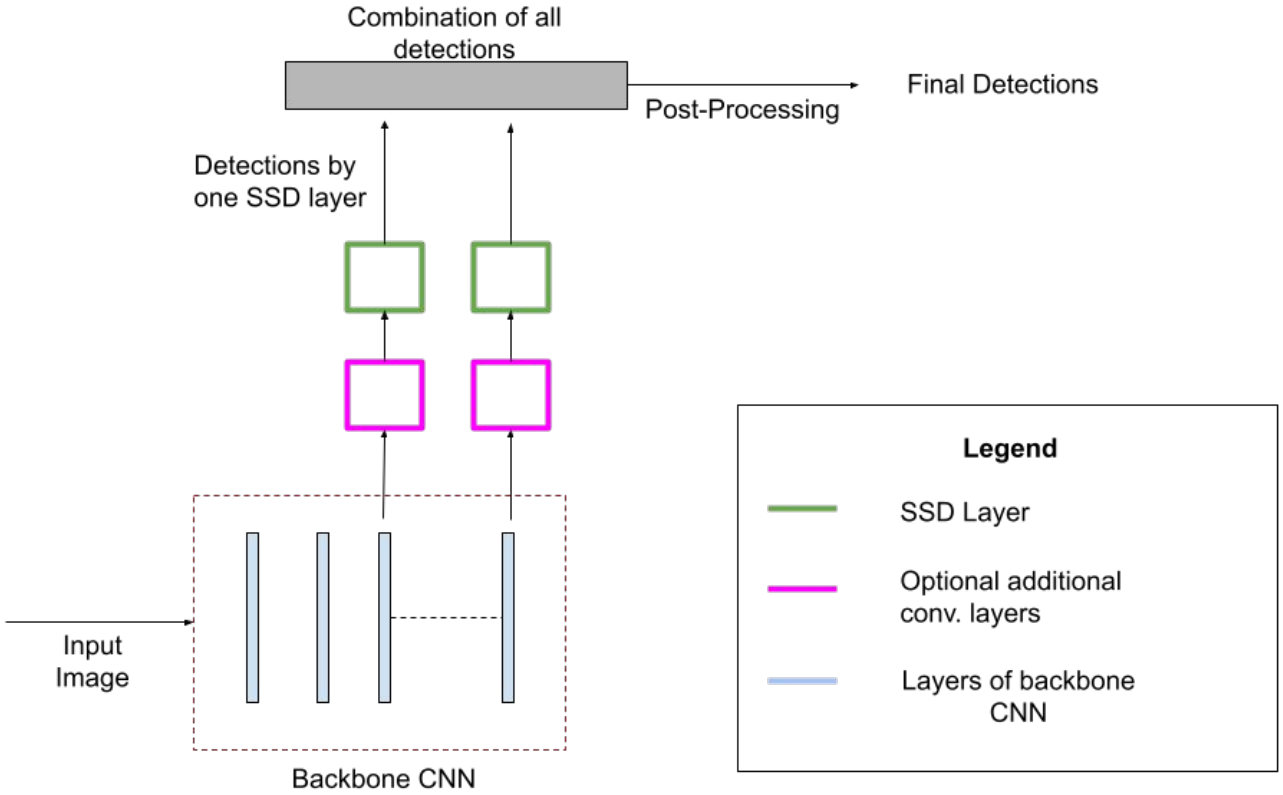
Object Detection in Deep Learning-2

Ujjwal

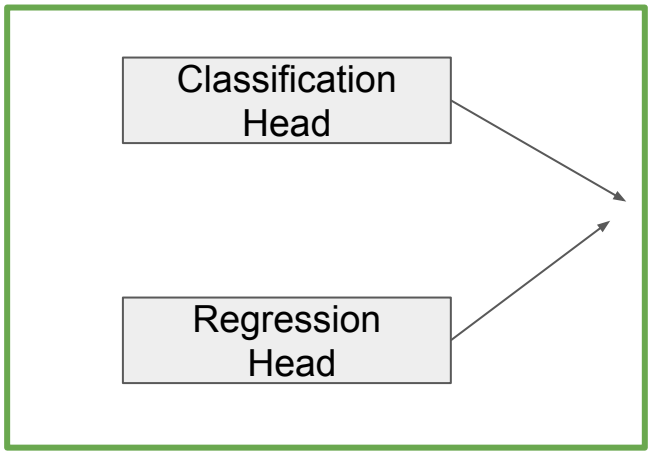
Previous Lecture

- Object Detection Evaluation
  - IoU
  - Precision and Recall
  - Precision Recall Curve
- Faster-RCNN
  - RPN
  - Loss functions for RPN and RCNN
  - Focal Loss
  - Smooth-L1 Loss

Single-Shot multiBox Detector (SSD)

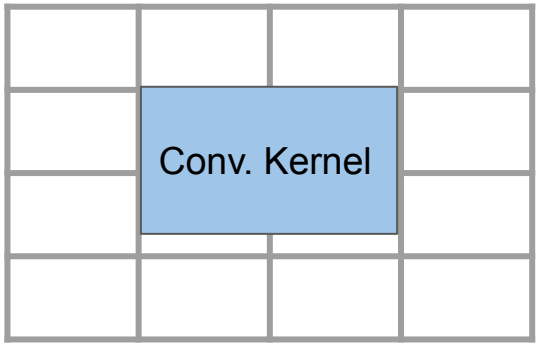


SSD Layer

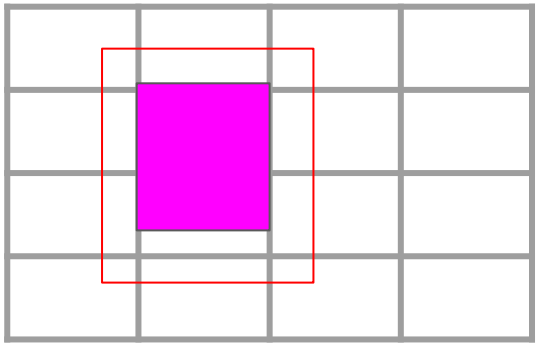


1. The classification and regression heads perform the same operation as in Faster-RCNN.
2. However, in SSD these heads operate like RPN.

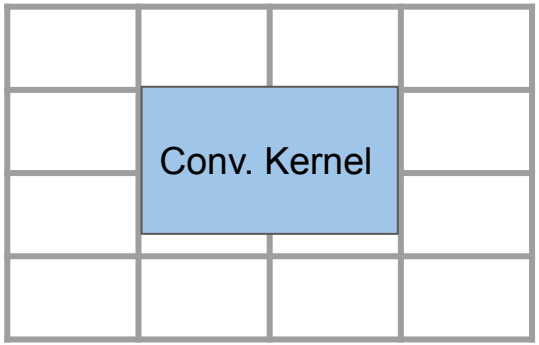
RPN vs. RCNN vs. SSD : Working on heads



1. A convolutional kernel operates over a feature map.
2. The resulting feature is used for classification or bounding box regression



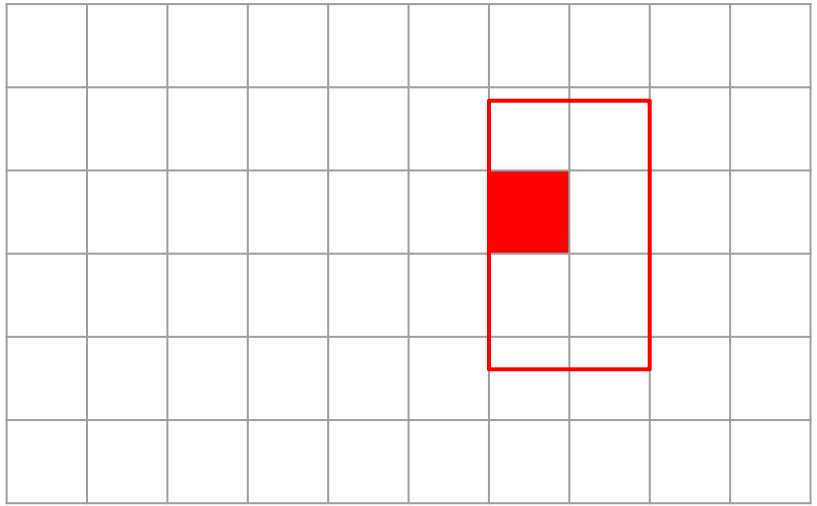
1. A RPN proposal is cropped from the feature map and resized to a fixed size
2. The resized feature is used for classification or bounding box regression.



1. Same as in RPN.
2. However RPN performs binary classification.
3. In SSD the classification is (N+1) for N-class object detection.

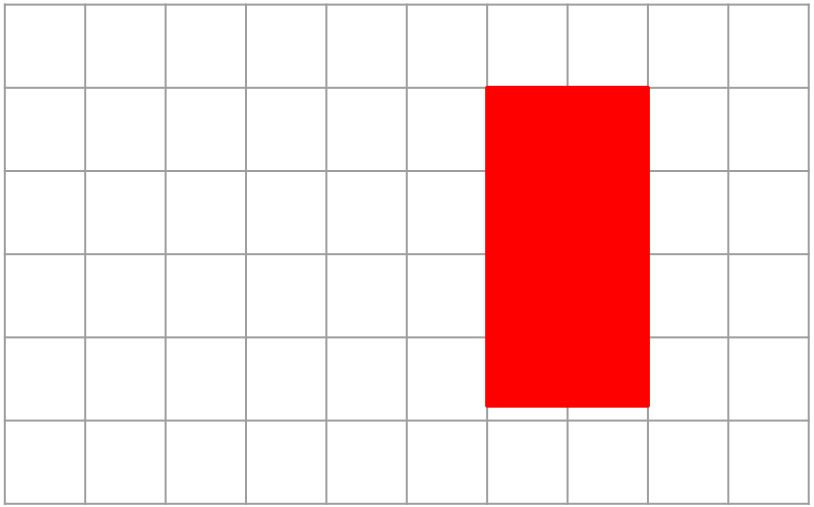
Need for RCNN

Image



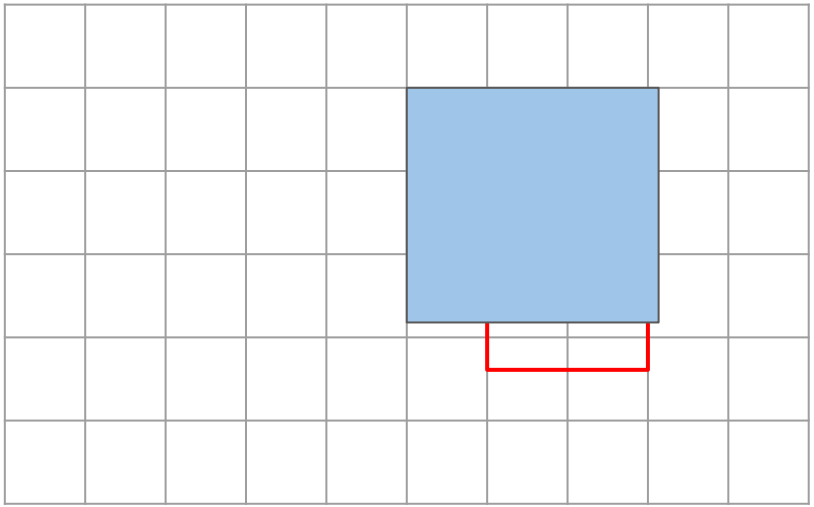
Feature Map with RPN proposal predicted at the shaded location

Need for RCNN



1. RCNN will crop the region inside the proposal and use it for prediction.

2. This covers a better profile of the object as seen in the example.



1. RPN will use a convolutional layer centered at that location (3x3 in this illustration).
2. This will incompletely cover the profile of the object.

Why not directly use RCNN ?

1. If RPN is not used :
    a. There will be no proposal to pool from.
    b. The RPN provides an initial rough estimate.

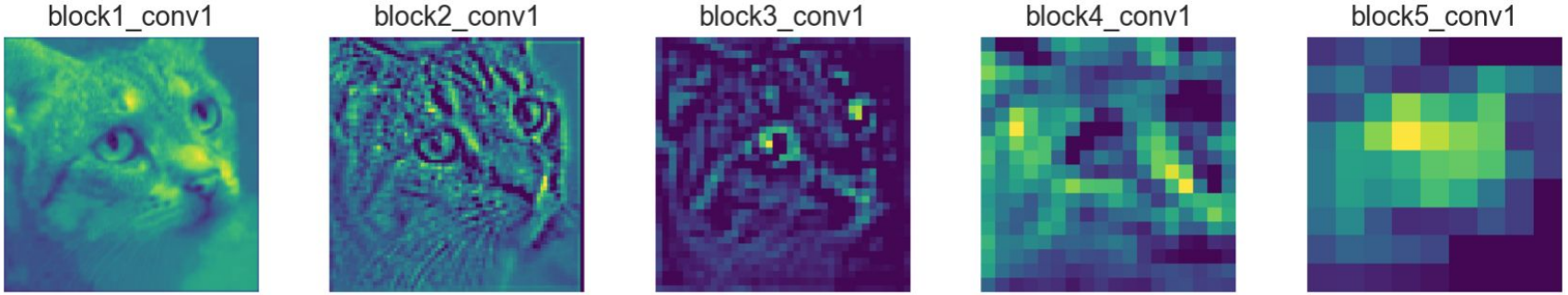
2. The rough estimate of the bounding box is used by RCNN to extract relevant features and provide a high quality detection.

SSD Layer

- Similar to Faster-RCNN in SSD as well :
  - The classification head has 2R filters where R is number of anchors at each location.
  - The regression head has 4R filters where R is number of anchors at each location.
- The anchors are used in the same was as in Faster-RCNN with one minor change:
  - There is a single threshold based on which a positive or negative anchor is selected.
- The SSD does (N+1)-class classification for a N-class object detection problem.

SSD is multi-scale

- SSD performs detections from multiple convolutional layers of the backbone.
- This allows it to take advantage of feature diversity across various convolutional layers



Visualization of features of an object through various layers of ResNet-50

SSD Loss Function

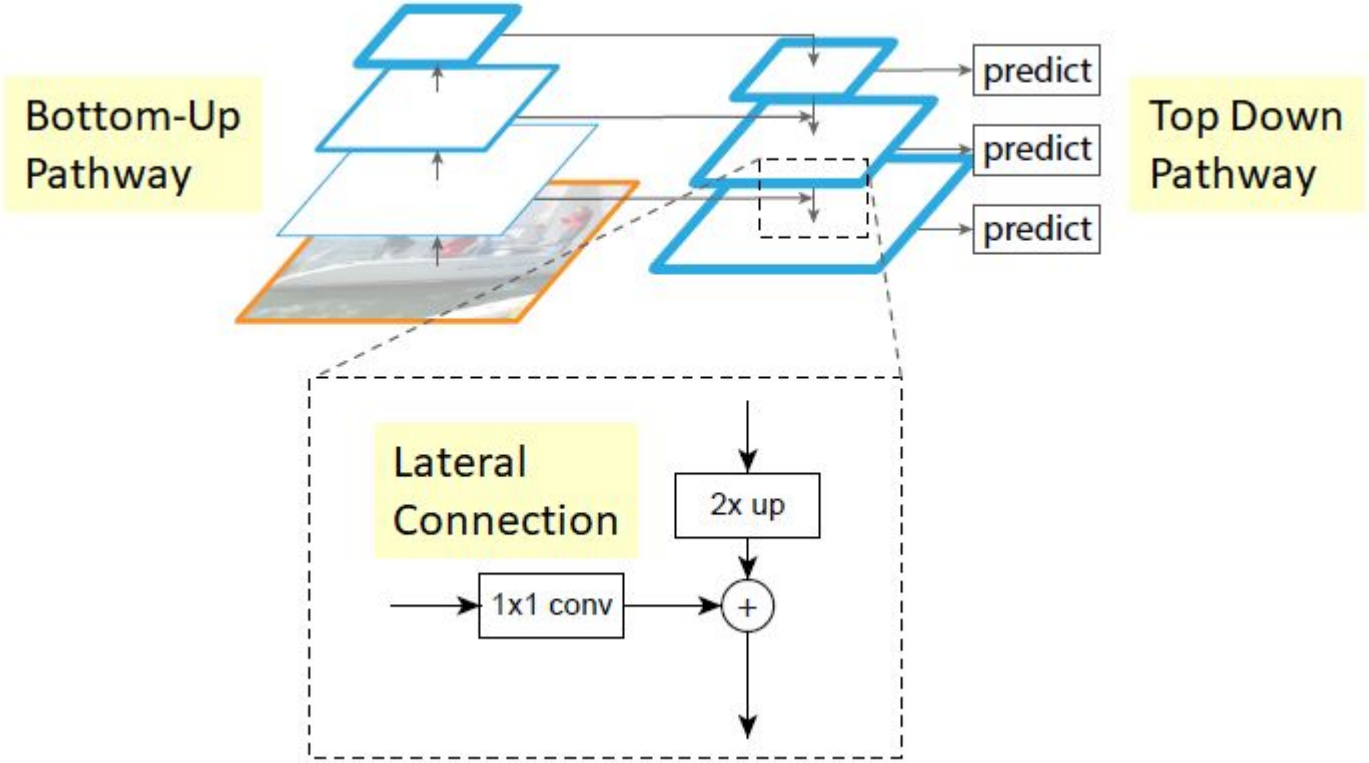
SSD Loss function is exactly the same as RPN loss function:

- There are N+1 classes instead of 2.
- In SSD literature this loss function is also known as multibox loss function.

One-Stage vs. Two-Stage Detectors

- Faster-RCNN is a two-stage detector:
  - It first provides a set of proposals (Stage 1 : RPN)
  - It then refines the proposals to get final detections (Stage 2 : RCNN)
  - Due to two stages and feature pooling it is slower but more accurate
- SSD is a one-stage detector:
  - It directly provides detections.
  - Due to one stage and simply convolutions it is faster but less accurate.

Feature Pyramid Networks (FPNs)



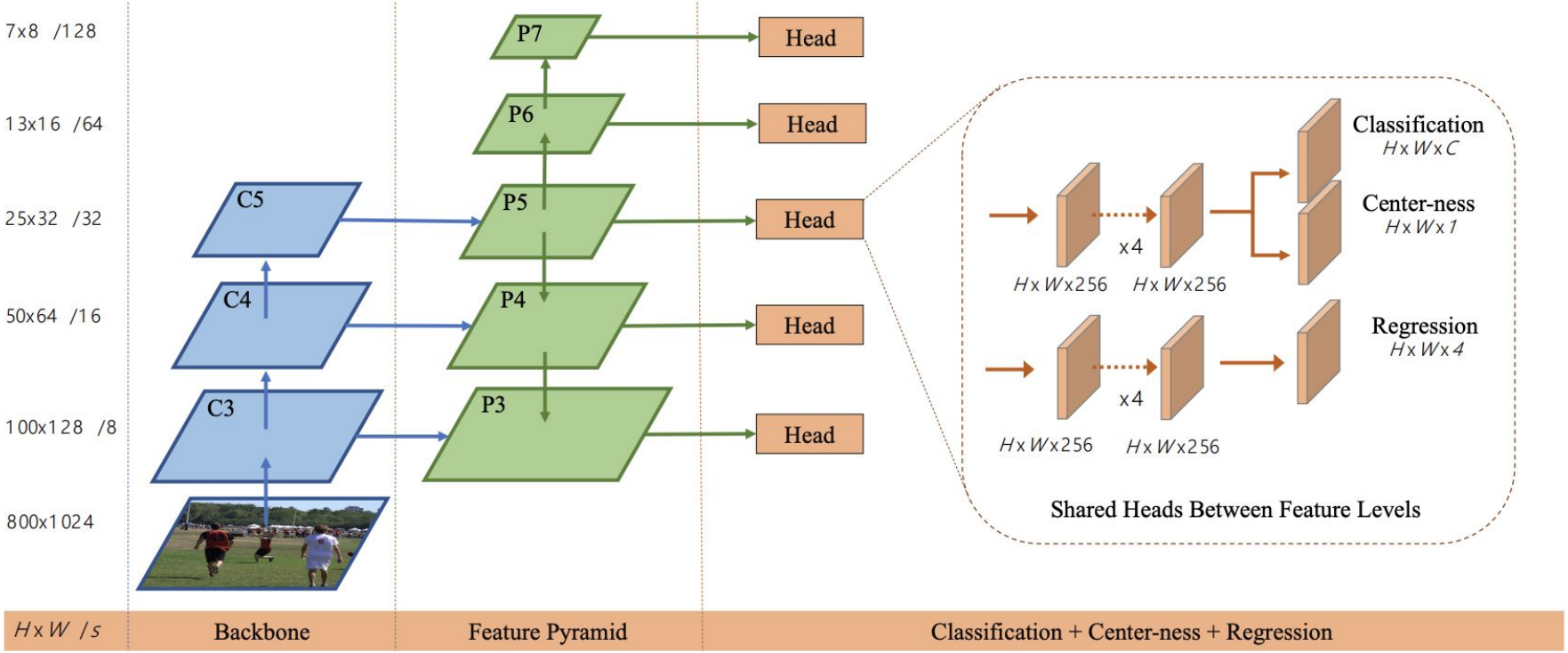
Salient aspects of FPN

- FPN allows multiple layers of a CNN to interact laterally (horizontally !!) in addition to vertically.

- This allows an extra way of interaction which has been shown to learn better features.

- FPN is a framework which can be adopted by both one-stage and two-stage detectors.

FCOS Detector



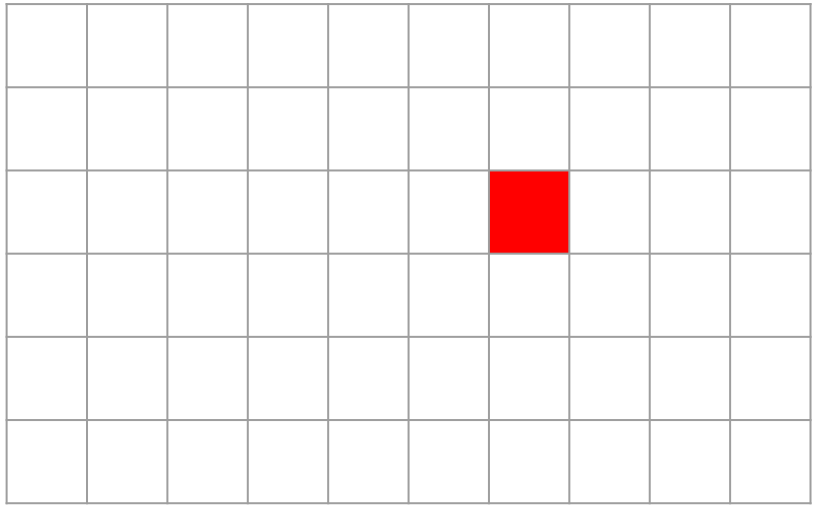
FCOS is Anchor free.

- FCOS does not use any anchors.
- This begs following questions:
    - Why anchors are necessary ?
    - How does FCOS get around the need to use anchors ?

Why Anchors are necessary ?

- Anchors provide a way to determine :
  - Positive and negative targets for computing loss function.
  - Overlap of an anchor with a groundtruth allows use to :
    - Decide the positive and negative targets when computing a loss.

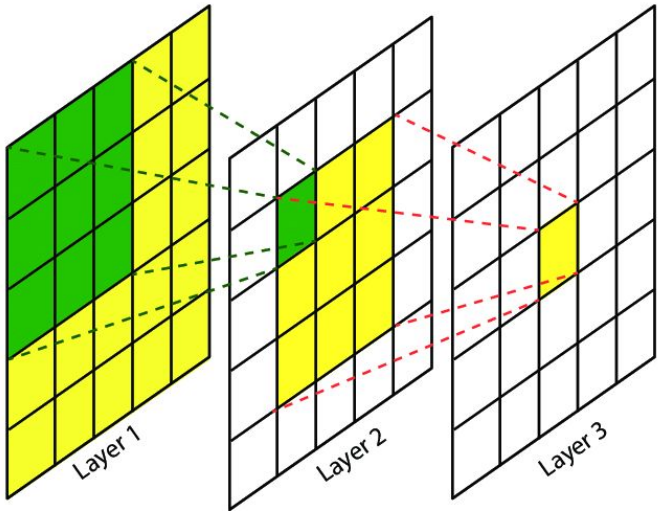
How FCOS ignores anchors ?



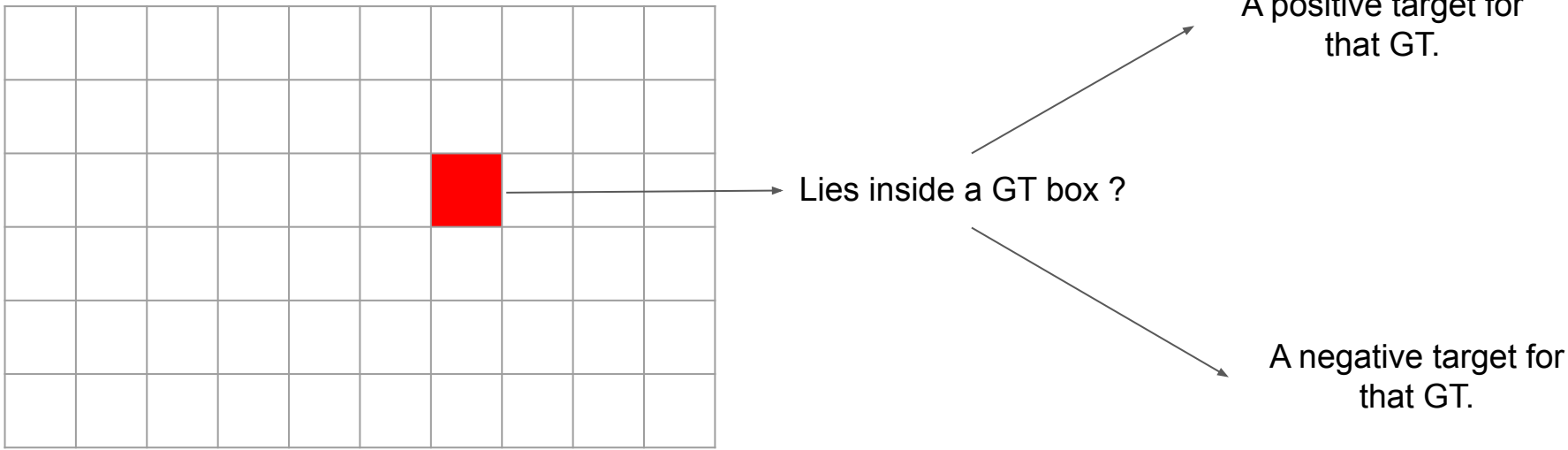
Does the marked cell in the feature map lie inside the bounding box of this person ?

Receptive Field of a feature map

- Let the input image size be M x N.
- Let the output feature map size be M/s x N/s
- Then one cell in the feature map corresponds to a s x s.
- This is known as receptive field of the output feature map.



FCOS Target assignment

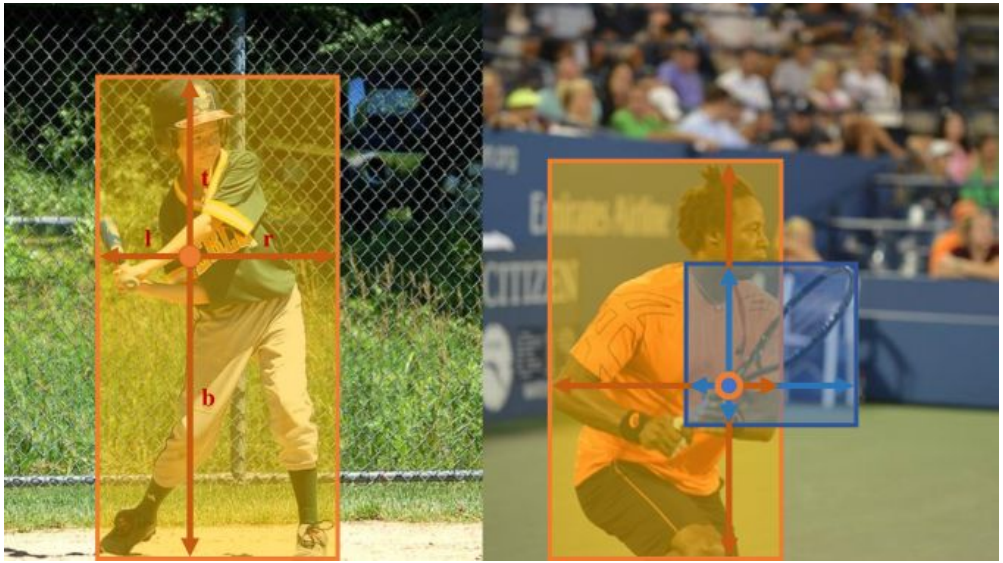


FCOS Target Assignment

For a positive target

- Classification is computed normally.
- Regression is directly predicting the bounding box offsets from that location.

FCOS Target Assignment



When a location resides in multiple bounding boxes, the smallest bounding box is used for classification and regression.

FCOS Centerness

$$\text{centerness}^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}.$$

- Centerness is maximum for the center of a bounding box.
- The idea is to generate a heatmap with maximal values at the centers of bounding boxes.

Loss Function

- Loss is same as SSD Loss with an added centerness loss.
- The centerness loss is a binary cross entropy loss