Yolo

Jawwad Shamsi and Rizwan Qureshi

April 17, 2021

1 Introduction

You Only Looked Once Suppose three objects are needed to be classified

1. pedestrian

2. car

3. motorcycle

The 4th class is a background class.

Recall Sliding window for object detection and localization

$$y = \begin{pmatrix} P_c \\ b_x \\ b_y \\ b_w \\ b_h \\ C_1 \\ C_2 \\ C_3 \end{pmatrix} \tag{1}$$

1.1 Challenges

1. What if there are multiple objects in a sliding window?

2. What if an object spans over multiples windows?

3. What if the object is not a balanced rectangle?

2 Solution:Bounding Box Prediction

1. Divide an image into cells. For instance, 3x3 grid or 19x19 grid.

2. Apply convolution through the sliding window technique.

3. get the output for each grid. i.e., equation **??**for each cell.

1

4. upper left corner for each cell would be 0,0 and lower right would be 1,1

5. if an object is in multiple cells then take the cell with center of the object

6. if there are a large number of cells then chances for more than one object in a cell a re low

7. b_x and b_y denotes center point of the object within the cell. They will always be between 0 and 1

8. b_w and b_h denotes width and height of the box. Can they be greater than 1?

9. What if an item spans multiple boxes ?

3 Intersection Over Union

IoU (Intersection over Union) measures the accuracy of bounding box prediction. Figure 1 illustrates the difference between predicted vs actual.
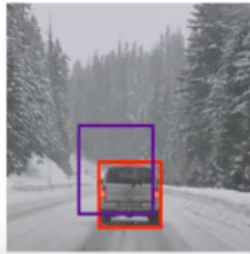

Figure 1: Predicted vs actual output

iou=size of intersection / size of union
(2)

if iou ==1 then same
if iou¿=0.5 then correct
iou is measure of overlap between two bounding boxes.

4   Non max Suppression

Developed to cater the problem of detecting same object multiple times It is possible when we have many cells. Each cell may likely claim to have an object (high probability)

A simplified car detection algorithm

$$y = \begin{pmatrix} P_c \\ b_x \\ b_y \\ b_w \\ b_h \end{pmatrix} \tag{3}$$

1. Discard all boxes with $P_c <= 0.6$

2. If there are more boxes

3. Pick the box with largest P_c prediction. Output it as a prediction

4. Discard any remaining box with $IoU >= 0.5$ with the output box in the above mentioned step

5. iterate

5   Anchor box algorithm

Designed to detect multiple objects with in a cell

Two anchor boxes. Each reflecting shape of the object Equation is modified

$$y = \begin{pmatrix} P_c \\ b_x \\ b_y \\ b_w \\ b_h \\ C_1 \\ C_2 \\ C_3 \\ P_c \\ b_x \\ b_y \\ b_w \\ b_h \\ C_1 \\ C_2 \\ C_3 \end{pmatrix} \tag{4}$$

5.1 Previously

Each object in training image is assigned to the grid cell that contains the object's midpoint.

Output y 3x3x8 (Why?)

5.2 Updated - with anchor boxes

Each object in training image is assigned to grid cell that contains object's midpoint and anchor box for the grid cell with the highest IoU.

Objects are assigned to a pair (gridcell,anchorbox) This instance . output = 3x3x2x8

5.3 limitations

cannot detect following:

1. two anchor boxes and 3 objects with in a cell

2. two objects with same anchor box shape with in a cell

These limitations can be addressed if we use a large grid size such as 19x19

5.4 how to select anchor boxes

1. Select by hand 8 to 10 most common shapes, which tend to cover most of the shapes

2. Use unsupervised machine learning (such as K-means) to identify most common shapes within the dataset

How to select size What is the smallest size box I want to detect? If center of bounding box differs with center of anchor box, we may miss the object detection.

5.5 How to detect objects using anchor boxes

1. Create several (may be hundreds) of anchor boxes

2. For each anchor box calculate which objects bounding box has highest overlap w.r.t IoU.

3. If IoU >=50 percent detect the object with highest IoU

4. else if IoU percent detection is ambiguous

5. else if <40 percent ; no object

6 Yolo-combined

Suppose two anchor boxes. THree classes 4th class is for background. y=3x3x2x8

1. For each cell get 2 predicted bounding boxes.

2. Remove low probability predictions

3. for each class (P,C,M) use non-max suppression to generate final predictions

7 R-CNN

There are four attributes, which are used to form an object
   scales, colors, textures, and enclosure
   Regions are identified based on these attributes.
   sub segments are generated to identify multiple regions
   Similar regions are then combined to form large regions
   Regions of Interest (RoI) are identified.