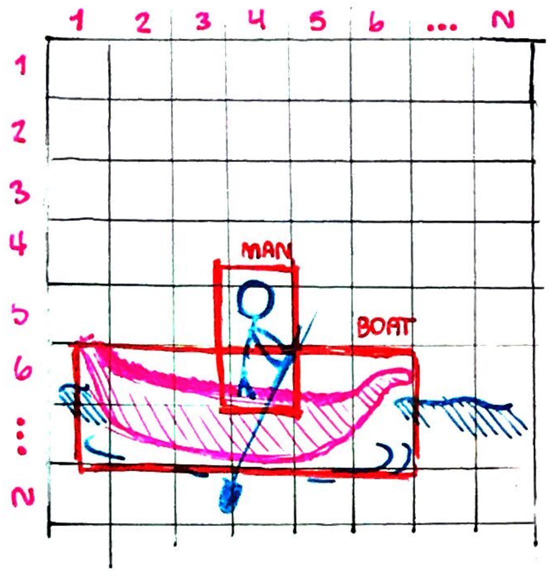
YOLO (YOU ONLY LOOK ONCE)



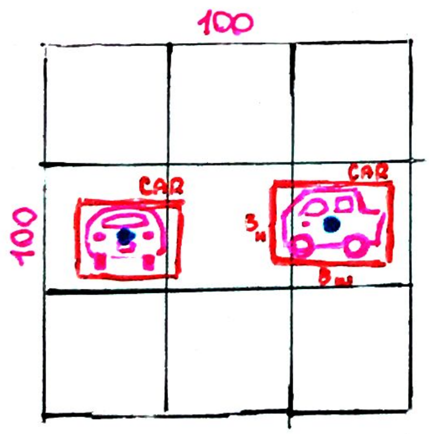
N x N CELLS / IMAGE

"ANCHOR BOXES" (□)

- REQUIRES 2 PARAMETERS:
  - N. OF BOUNDING BOX PREDICTIONS PER CELL
    - + MORE PREDICTIONS → MORE OBJECTS THAT CAN BE DETECTED IN CLOSE PROXIMITY.
    - + MORE PREDICTIONS → MORE COMPUTING POWER NEEDED / MORE TRAINING DATA NEEDED.
  - SHAPE OF THE BOUNDING BOXES
    - + EVERY BOX HAS A DIFFERENT SPECIALIZATION. FOR EXAMPLE:
      - 1ST BOX: SMALL FLAT RECTANGLE ▭
      - 2ND BOX: SMALL TALL RECTANGLE ▯
    - + ITS A WASTE TO MAKE A BOX SPECIALIZE IN SHAPES THAT RARELY APPEAR IN DATA

- YOLO v2 PROPOSES THE K-MEANS CLUSTERING ALGORITHM TO DETERMINE THE IDEAL NUMBER AND SHAPE OF THE BOXES.

TRAINING LABELS

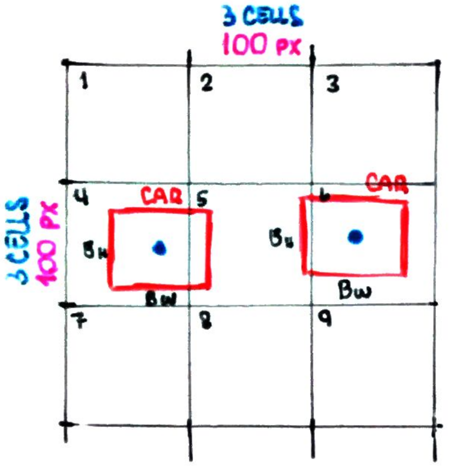


100 x 100 PIXELS
3 x 3 CELLS

- PER GRID CELL YOU'LL HAVE A VECTOR:

$$y = \begin{bmatrix} P_c \\ B_x \\ B_y \\ B_H \\ B_W \\ C_1 \\ C_2 \\ \vdots \\ C_N \end{bmatrix}$$

- P_c: PROBABILITY OF CENTER OF OBJECT PRESENT IN CELL? (0/1)
- B_x: RELATIVE BOX X COORD (CENTRE POINT)
- B_y: RELATIVE BOX Y COORD (CENTER POINT)
- B_H: BOX HEIGHT
- B_W: BOX WIDTH
- $C_1, C_2, \ldots, C_N$: PROBABILITY. AS MANY C'S AS CLASSES TO DETECT
  - DURING TRAINING THIS USES ONE HOT ENCODING
  - DURING NORMAL USE THEIR VALUES SUM 1



EXAMPLES:

GIVEN CLASS 2 = "CAR"

GRID # 1

$$y = \begin{bmatrix} 0 \\ x \\ x \\ x \\ x \\ x \\ \vdots \\ x \end{bmatrix}$$

X = DONT CARE

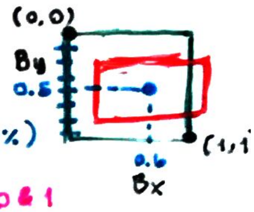
GRID # 4

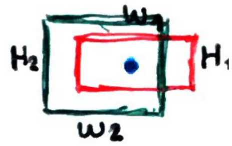
$$y = \begin{bmatrix} 1 \\ B_x \\ B_y \\ B_w \\ B_H \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

CAR PRESENT (100%)

0.6, 0.5 } ALWAYS BETWEEN 0 & 1

0.8, 0.95 } COULD BE > 1



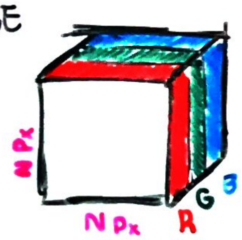


$B_W = W_1 / W_2$

$B_H = H_1 / H_2$

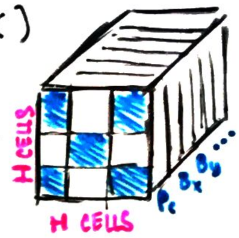
INPUT AND OUTPUT SHAPE

IMAGE



→ CONV → MAXPOOL ... →

TENSOR (MATRIX)



INPUT (4 DIMS)

$m \times N \times N \times 3$

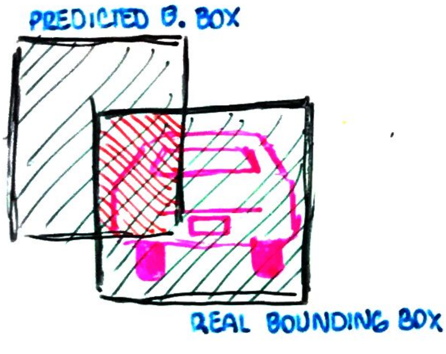
N. OF IMAGES, PIXELS, PIXELS, COLOR CHANNELS

OUTPUT (5 DIMS)

$m \times H \times H \times B \times (5 + N_c)$

N. OF IMAGES, CELLS, CELLS, No OF B. BOX PREDICTIONS PER CELL, N. OF CLASSES

INTERSECTION OVER UNION



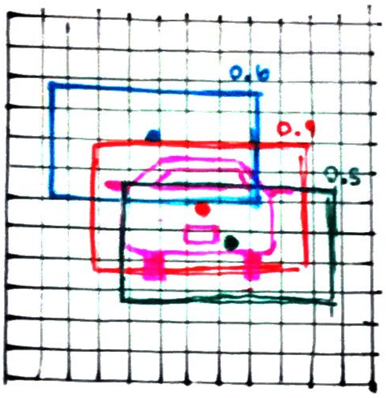
- INTERSECTION OVER UNION (IoU)

$$\frac{\text{SIZE OF RED REGION}}{\text{SIZE OF GREEN REGION}} \quad \begin{matrix}\text{(INTERSECTION)} \\ \text{(UNION)}\end{matrix}$$

- CONVENTION: IoU $\geq$ 0.5
  CORRECT PREDICTION

- THE CONVENTION CAN BE CHANGED TO BE MORE STRINGENT. IT'S NOT RECOMENDED TO DECREASE THE THRESHOLD.

NON-MAX SUPPRESSION (NMS)



12 x 12 CELLS

AFTER YOLO GENERATES ITS PREDICTIONS FOR BOUNDING BOXES, MANY OF THE PREDICTIONS COULD BE DUPLICATES OR LESS ACCURATE DETECTIONS OF THE SAME OBJECT. TO "CLEAN" THE DETECTIONS AN ALGORITHM CALLED "NON-MAX SUPPRESSION" IS USED. IT CONSISTS OF TWO PARTS.

PRE-NMS FILTER:

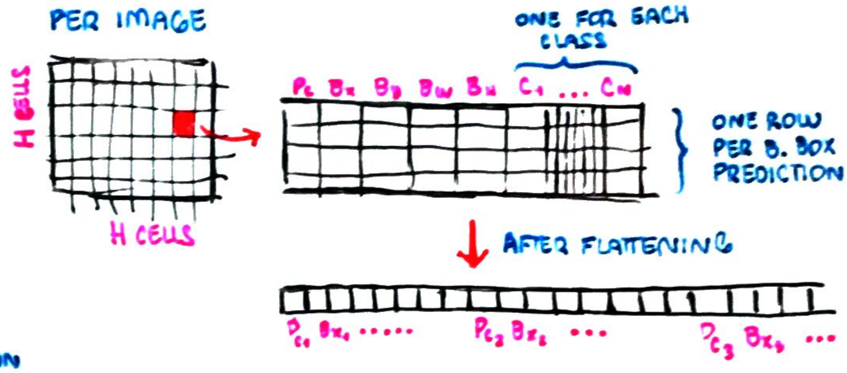
- THE LAST TWO DIMENSIONS OF THE OUTPUT TENSOR ARE FLATTENED

OUTPUT DIMS

$m \times H \times H \times B \times (5 + N_c)$

AFTER FLATTENING

$m \times H \times H \times \underbrace{(B \times (5 + N_c))}_{\text{SINGLE DIMENSION}}$



- CALCULATE THE "BOX SCORES" BY MULTIPLYING THE PRESCENCE PROBABILITY (P_c) OF EACH B. BOX WITH ITS CORRESPONDING VECTOR OF CLASS PROBABILITIES ($C_1, C_2, C_3 \dots C_N$)

$$\text{SCORES} = P_c * \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ \vdots \\ C_N \end{bmatrix} = \begin{bmatrix} P_c C_1 \\ P_c C_2 \\ P_c C_3 \\ \vdots \\ P_c C_N \end{bmatrix}$$

ELEMENT-WISE MULTIPLICATION

NEW OUTPUT DIMS

$m \times H \times H \times (B \times 5)$

- EXTRACT THE INDEX (ARGMAX) AND VALUE (MAX) OF THE HIGEST SCORE. THIS RESULTS IN 1 CLASS PREDICTION PER B. BOX PREDICTION

- FILTER ALL SCORES BY COMPARING TO A THRESHOLD AND CREATING A MASK.

$$[0.5 \;\; 0.9 \;\; 0.7 \dots 0.1] \leq 0.6$$
$$[\text{FALSE} \;\; \text{TRUE} \;\; \text{TRUE} \dots \text{FALSE}]$$

NMS FILTER :

FOR EACH CLASS

1. TAKE THE BOX WITH THE HIGHEST P_c
2. GET THE IoU WITH ALL OTHER BOXES.
3. DELETE ANY BOX THAT HAS AN IoU VALUE HIGHER THAN A THRESHOLD (COMONLY 0.5).
4. SELECT THE NEXT BOX WITH THE HIGHEST P_c AND REPEAT STEP 2-3.
5. REPEAT UNTIL THERE AREN'T ANY REMAINING BOXES WITH A HIGH P_c