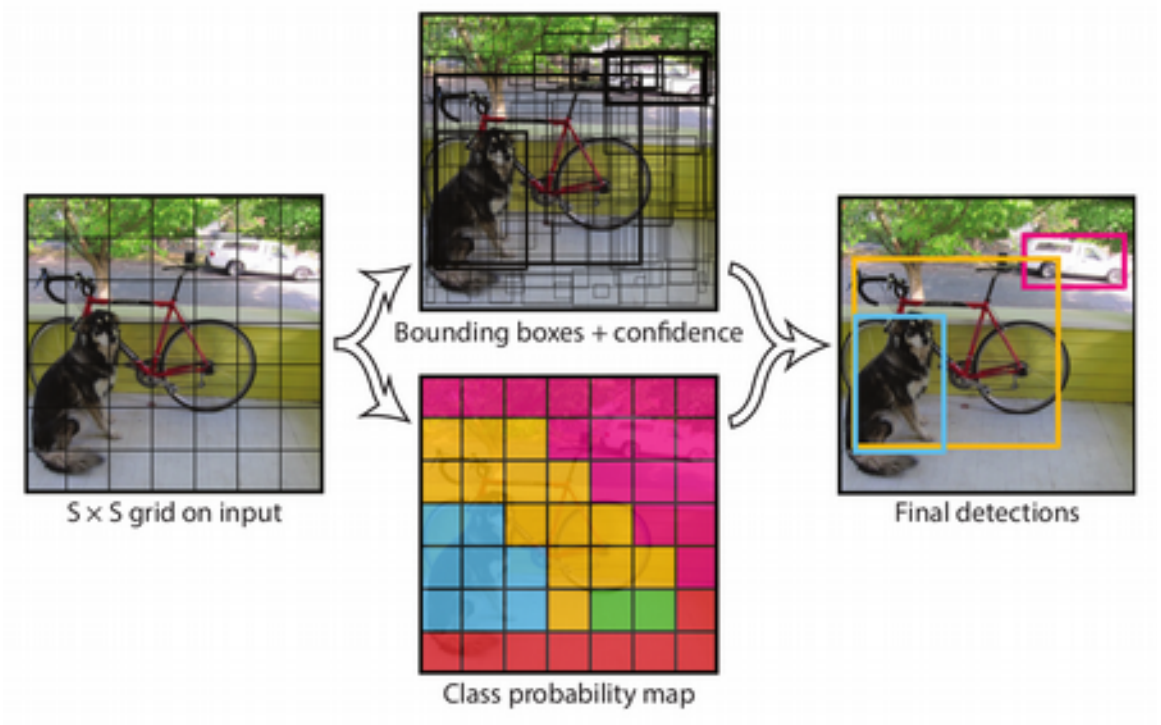
DETECCIÓN de OBJETOS
Parte II: Hacia un entorno Real-Time

José M. Saavedra R.

YOLO: You Only Look Once

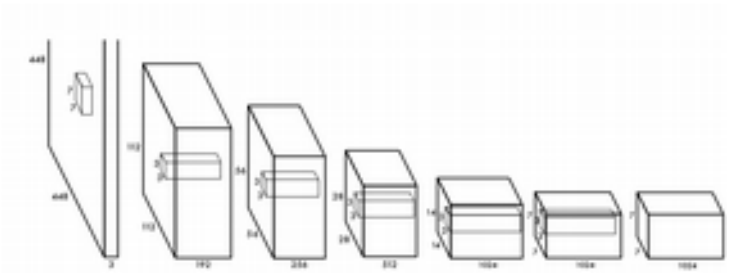
La diferencia clave con respecto a Faster RCNN es que tanto la detección de regiones como la clasificación se realiza en un solo paso.



YOLO: You Only Look Once

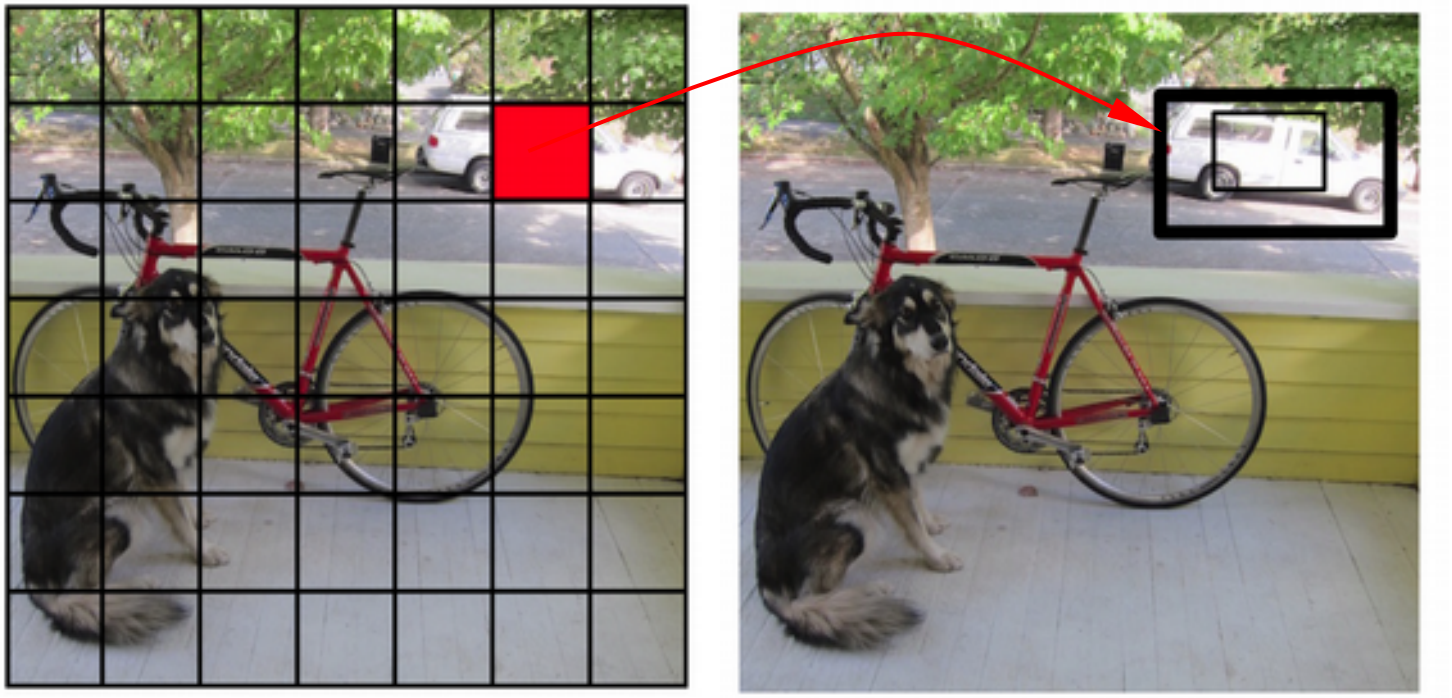
La diferencia clave con respecto a Faster RCNN es que tanto la detección de regiones como la clasificación se realiza en un solo paso.

Reduced Spatial Dimensionality (Split)
A través de capas convolucionales, similar a Faster-RCNN.
Una celda indicará el campo receptivo de una neurona en la capa de características.



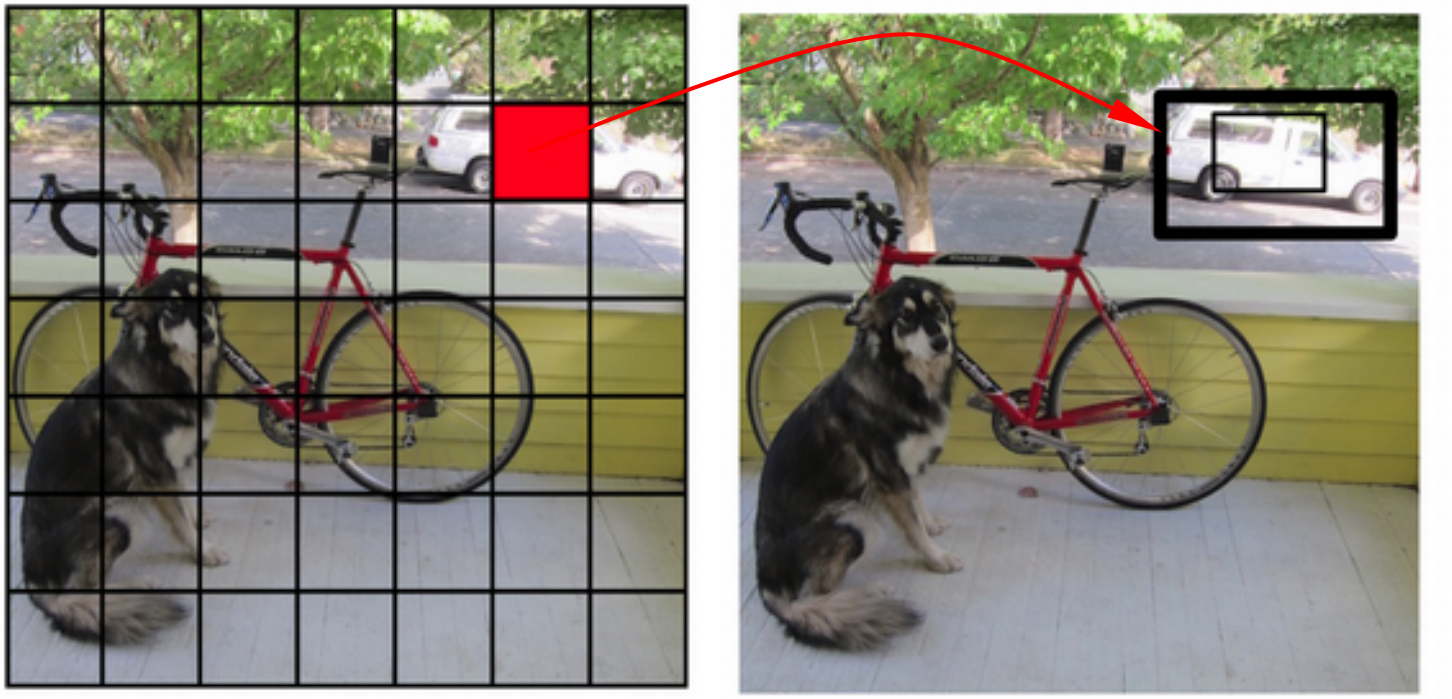
YOLO: You Only Look Once

Predicted Boxes
Cada celda se encargará de predecir B “bounding boxes” + un valor de confianza



YOLO: You Only Look Once

Predicted Boxes
Cada celda se encargará de predecir B “bounding boxes” + un valor de confianza

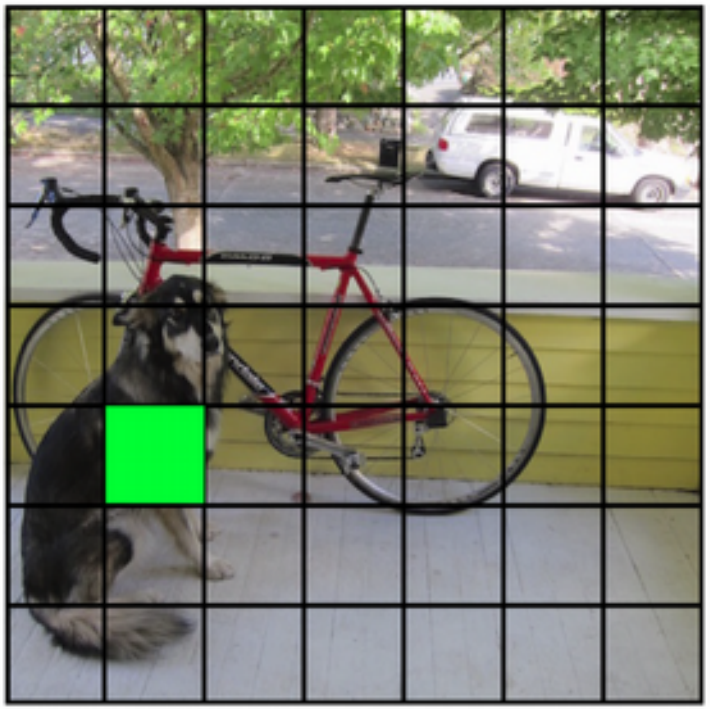


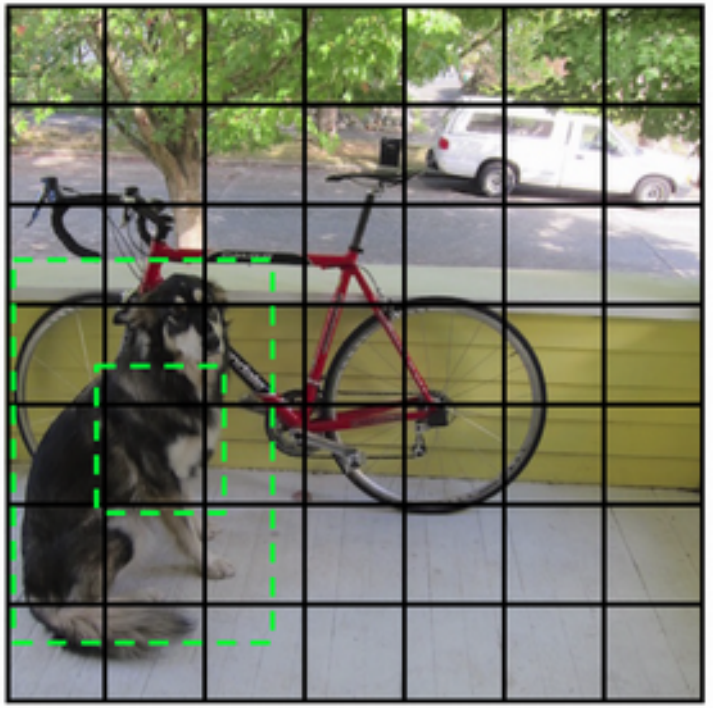
Box: x y w h + confidence [5 parámetros]

x, y: centro de la celda, con respecto al borde de la imagen.
w, h: ancho y alto relativo al tamaño de la imagen

YOLO: You Only Look Once

Predicted Boxes
Cada celda se encargará de predecir B “bounding boxes” + un valor de confianza





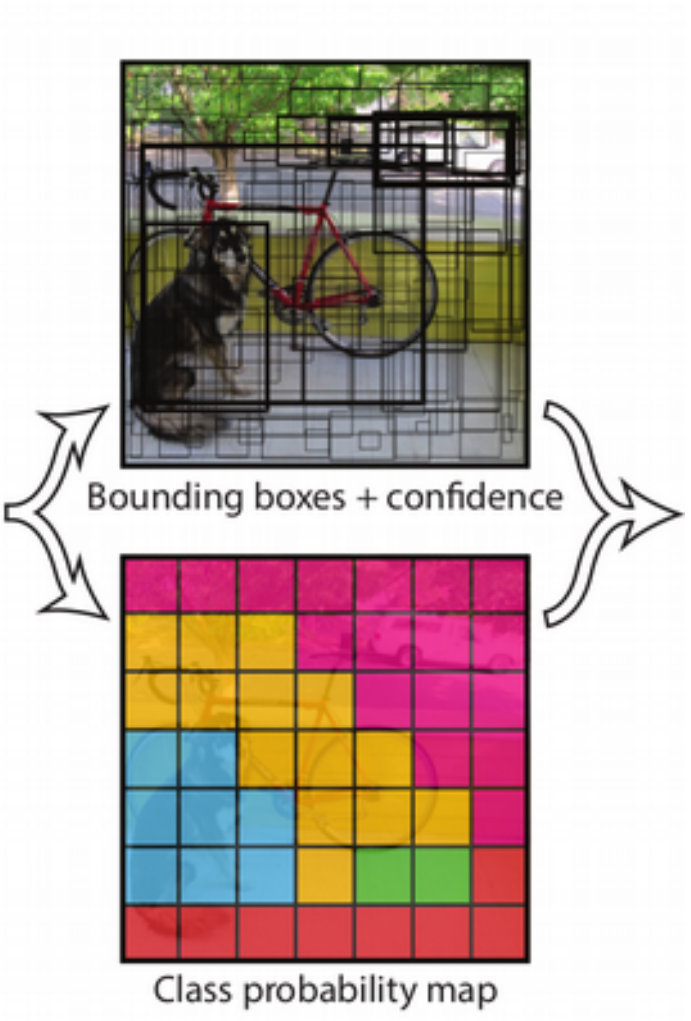
YOLO: You Only Look Once

Predicted Boxes
Cada celda se encargará de predecir B “bounding boxes” + un valor de confianza

¿Cómo manejar diferentes escalas?

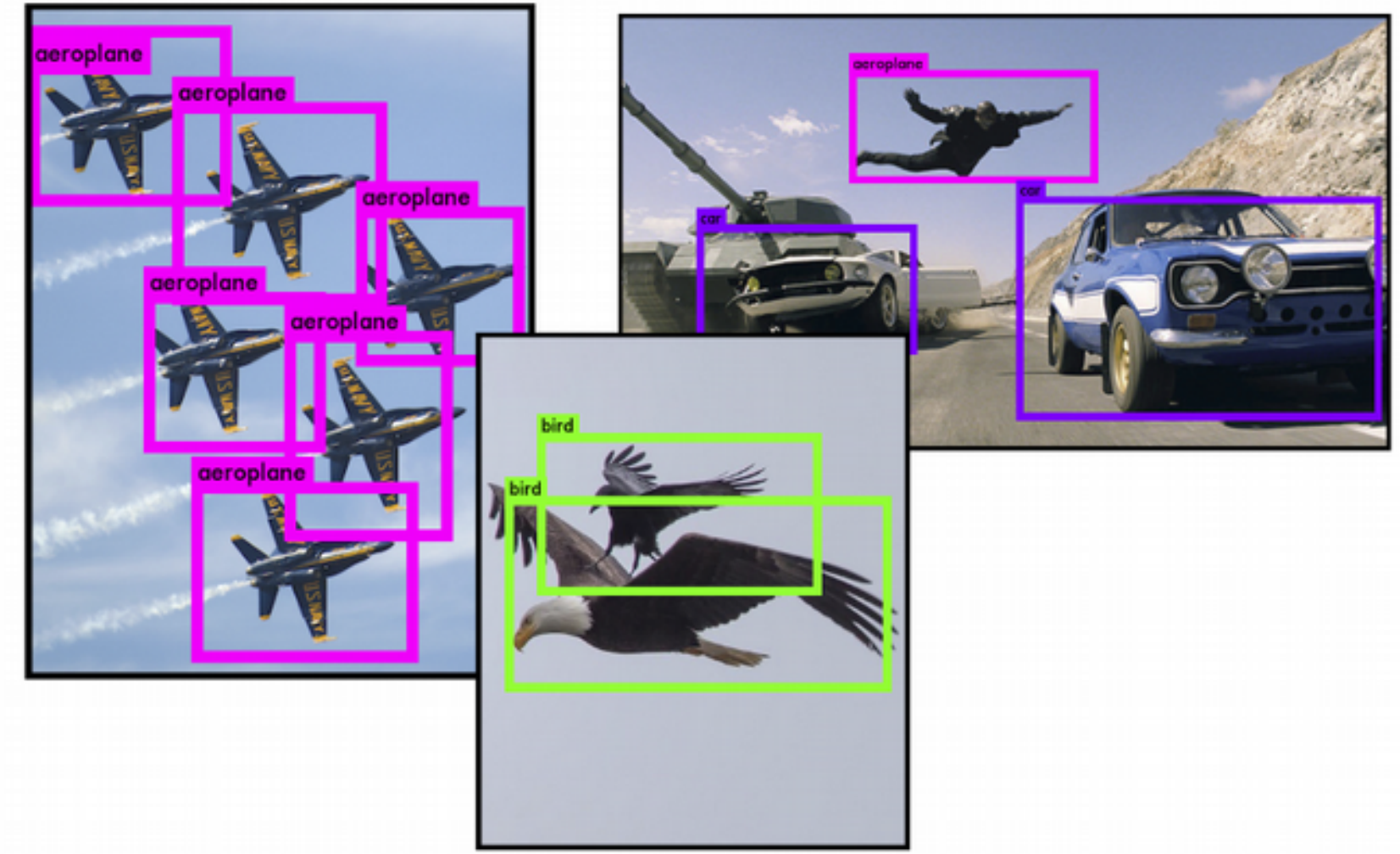
YOLO: You Only Look Once

Predicted Classes
Cada celda, además, está asociada a la probabilidad de ocurrencia de una clase dado que existe un objeto.



YOLO: You Only Look Once

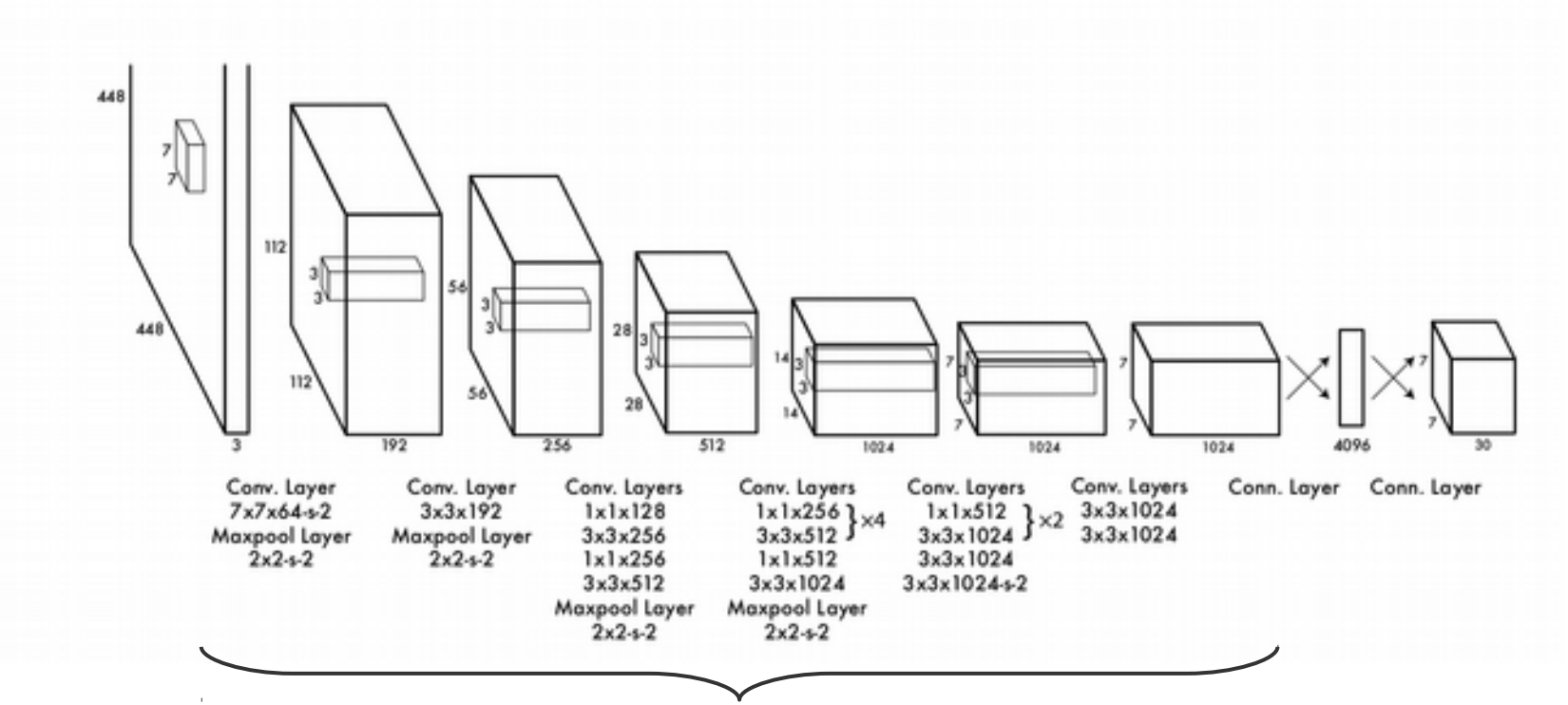
Final Detection



aavedra

YOLO: You Only Look Once

Modelo Convolucional



Pre-entrenado en ImageNet

YOLO: You Only Look Once

Loss

$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

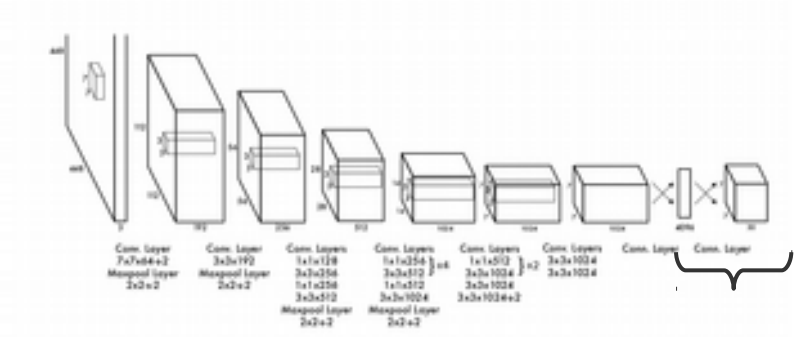
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

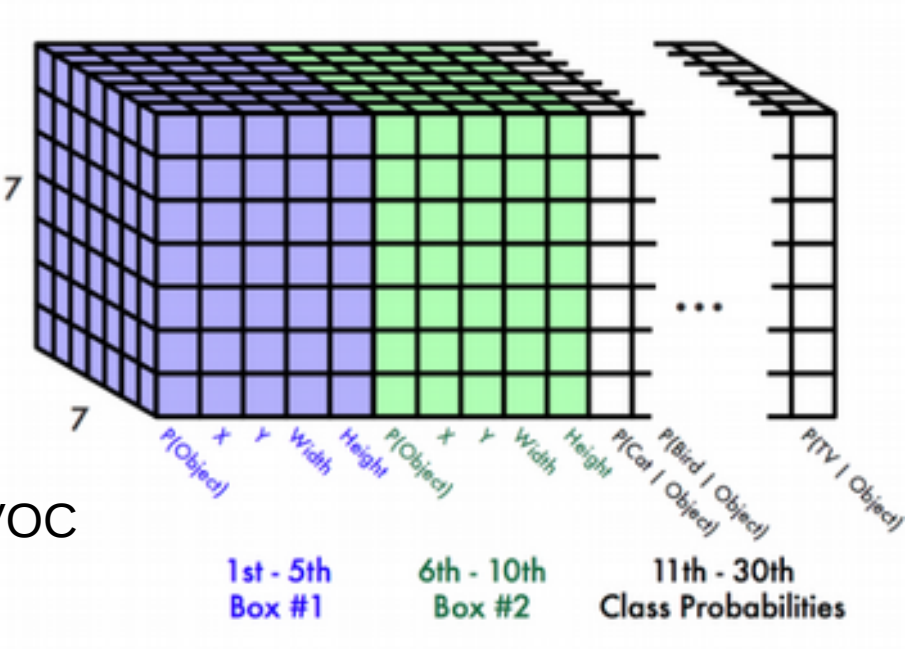
$$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2$$

$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2$$

YOLO: You Only Look Once

Modelo de Regresión





Salida: regresor de (SxSx(B*5+C))

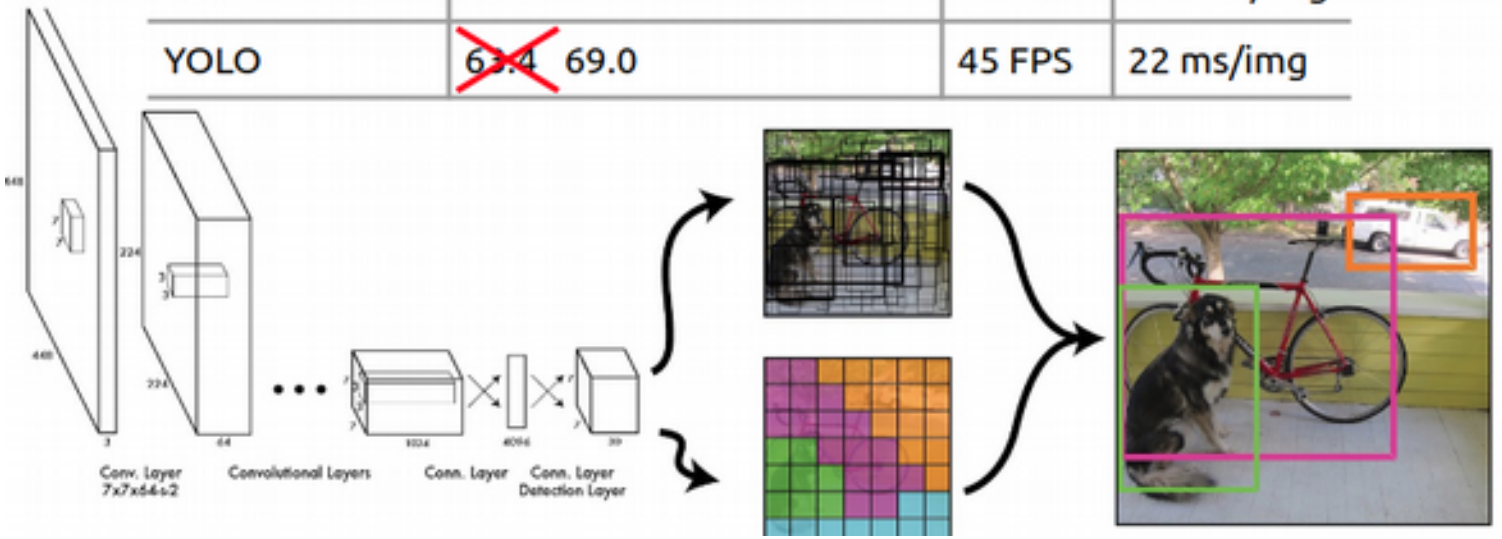
S = 7
B = 2
C = número de clases, e.g. 20 en Pascal VOC

Así, salida → 7x7x30

YOLO: You Only Look Once

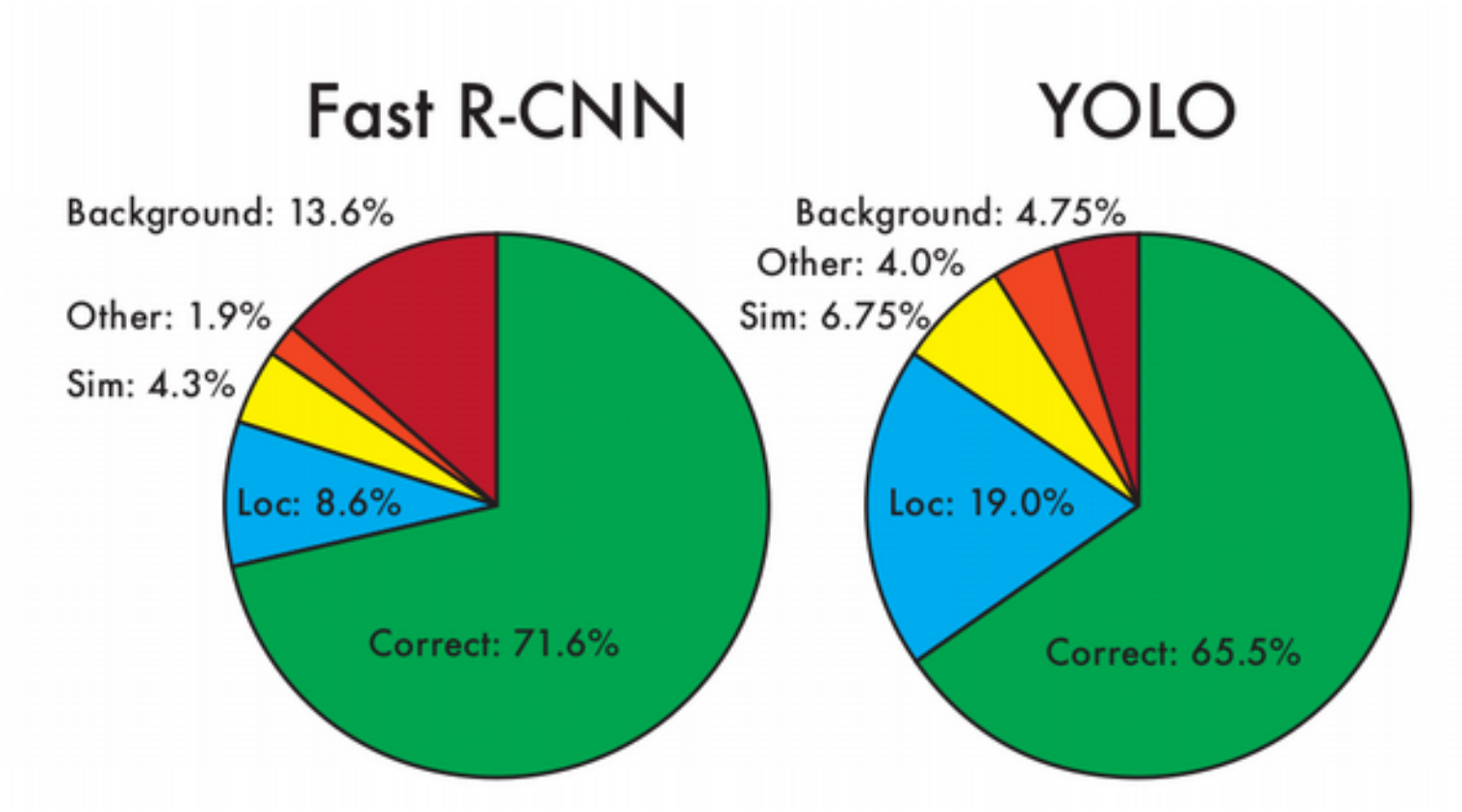
Desempeño

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img
YOLO	~~63.4~~ 69.0	45 FPS	22 ms/img



YOLO: You Only Look Once

Menos falsos positivos en background, pero mala localización



YOLO 9000

Usando lo mejor de Faster RCNN

Anchors

YOLO9000

Algunas mejoras

- Batch Normalization
- High Resolution Classifier (224 → 448, S = 13)
- **Anchor for Bounding Boxes**
- Número de anchors se definen a través de clustering.

YOLO9000

Anchors

Faster RCNN (Revisar clase Faster RCNN)

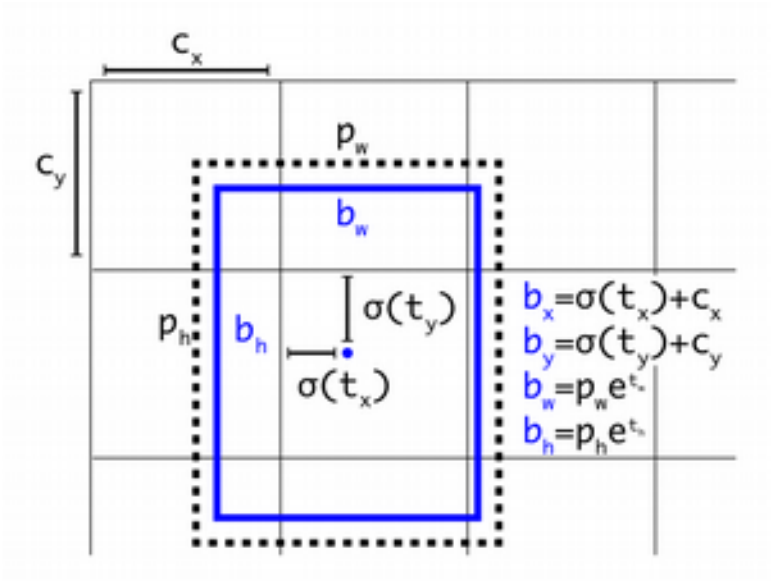
$$x = (t_x * w_a) - x_a$$
$$y = (t_y * h_a) - y_a$$

x,y no está limitado y puede caer en cualquier parte de la imagen generando inestabilidad

YOLO9000

Anchors

YOLO 9000 (x,y caen dentro de la celda correspondiente)



$$b_x = \sigma(t_x) + c_x$$
$$b_y = \sigma(t_y) + c_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$
$$Pr(\text{object}) * IOU(b, \text{object}) = \sigma(t_o)$$

sigma es la función logística, de modo que varíe entre 0 y 1

YOLO9000

Desempeño

Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19
YOLOv2 288 × 288	2007+2012	69.0	91
YOLOv2 352 × 352	2007+2012	73.7	81
YOLOv2 416 × 416	2007+2012	76.8	67
YOLOv2 480 × 480	2007+2012	77.8	59
YOLOv2 544 × 544	2007+2012	**78.6**	40

YOLOv3 [2018]