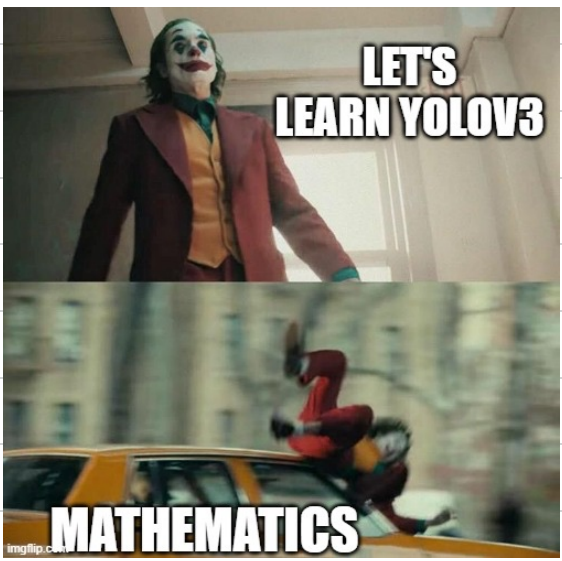
Session-8

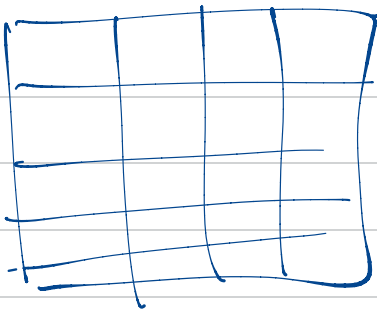
Object Segmentation

April 01, 2024



Agenda

1. Recap - Anchor boxes
2. Yolo - V3
3. Code walk through.
4. Segmentation - Transpose Convolution.
5. FCN
6. UNET



13 x 13 x 512

From VGG

(1) Single Anchor + Single Anchor box

Yolo →

13 x 13 x 512 → (1 x 1 x 512) x 8 → 13 x 13 x 8

13 x 13 x 8

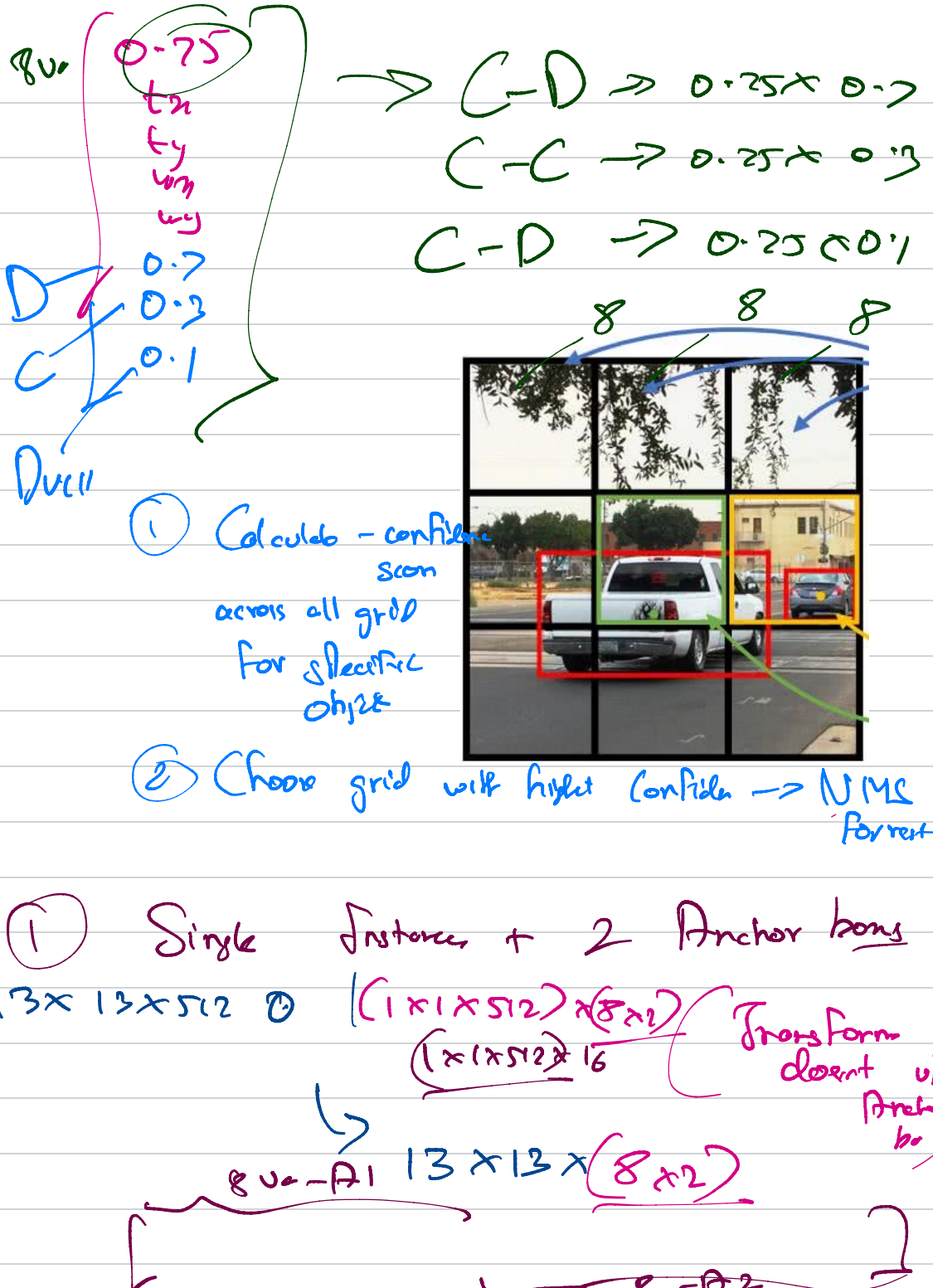
Objectness_Score + 1

4 - coordinat + 4

no of class) 3

3

multiply → Objectness x Class Prob



① Single Instance + 2 Anchor boxes

13 × 13 × 512 ⊙ (1 × 1 × 512) × (8 × 2) → (1 × 1 × 512 × 16) → Transform dimension up Anchor box

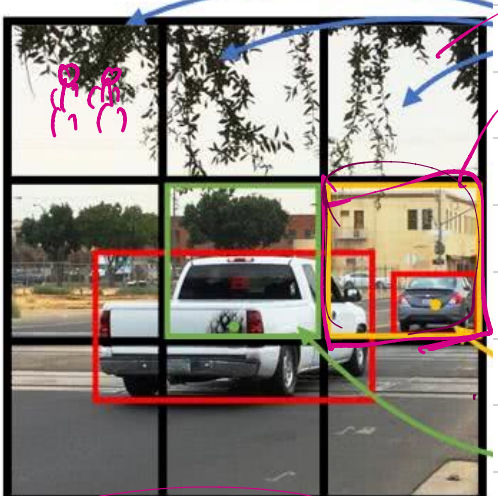
↳ 8v-A1 13 × 13 × (8 × 2)

8v-A2

1. Calculate confidence score across all grid for specific object

2. Choose grid with highest confidence → NMS for rest

3. Two Instance + Two Anchor box



→ 10 diff → Anchor box

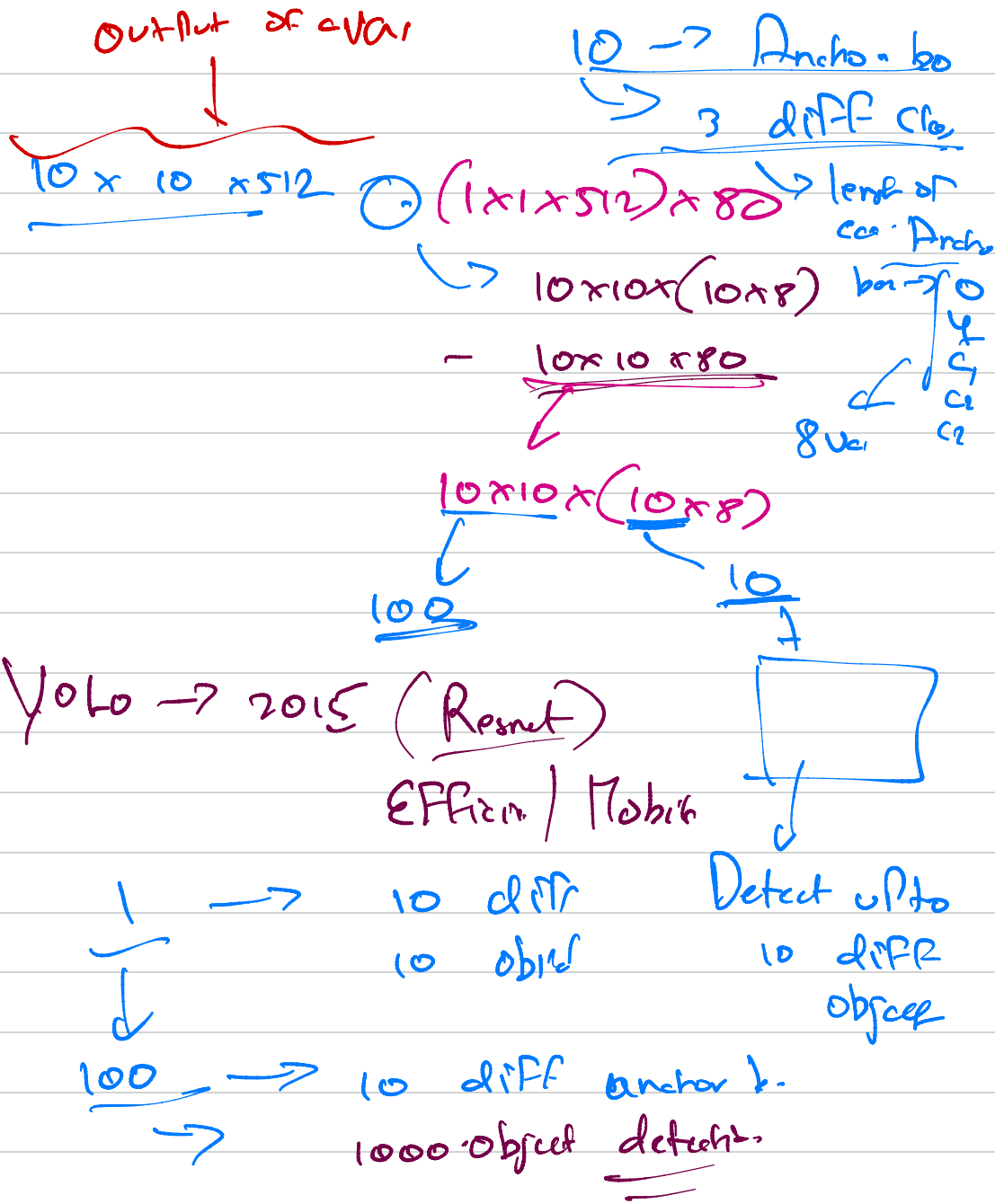
p_c b_x b_y b_w b_h c_1 c_2 c_3 — 8 - AL

p_c b_x b_y b_w b_h c_1 c_2 c_3 — 8 - AL

Loss function

→ 10 diff Anchor box

→ Grid - 10x10 diff no of object Scan detect


output of cnn
10 x 10 x 512
(1x1x512)x80
10 -> Anchor box
3 diff classes
length of coco Anchor box
10x10x(10x8)
- 10x10x80
4 coords
1 class
8 vals
10x10x(10x8)
100
10
Yolo -> 2015 (Resnet)
Effici / Mobile
1 -> 10 diff 10 objects
Detect upto 10 diff objects
100 -> 10 diff anchor b.
1000 object detect.

Yolo - V3

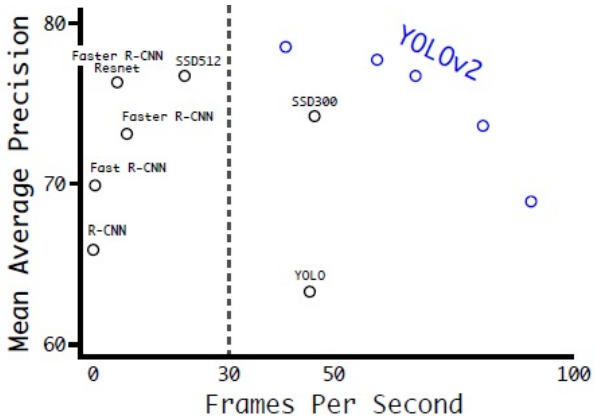
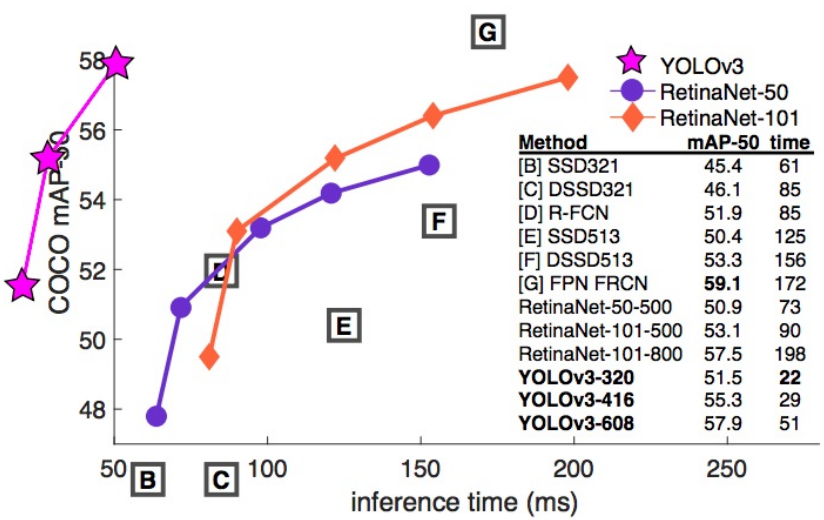


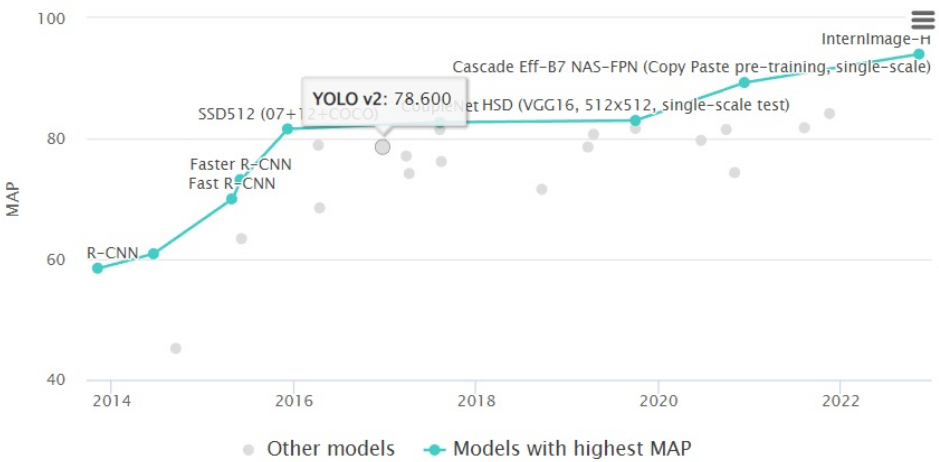
Figure 4: Accuracy and speed on VOC 2007.

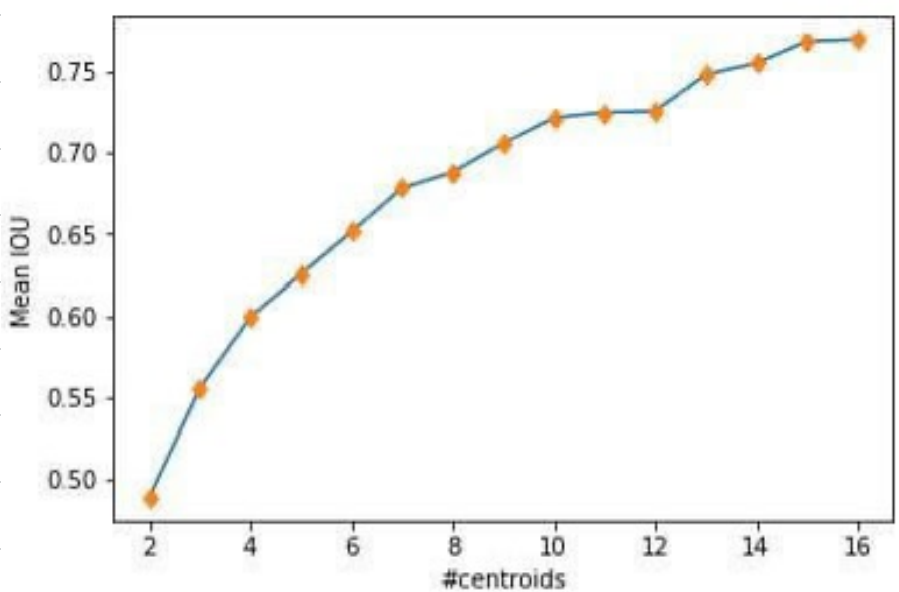


Method	mAP-50	time
[B] SSD321	45.4	61
[C] DSSD321	46.1	85
[D] R-FCN	51.9	85
[E] SSD513	50.4	125
[F] DSSD513	53.3	156
[G] FPN FRCN	**59.1**	172
RetinaNet-50-500	50.9	73
RetinaNet-101-500	53.1	90
RetinaNet-101-800	57.5	198
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

Where Are We Today

Object Detection on PASCAL VOC 2007





darknet -53

	Type	Filters	Size	Output
	Convolutional	32	3 x 3	416 x 416
	Convolutional	64	3 x 3 / 2	208 x 208
1 x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			208 x 208
	Convolutional	128	3 x 3 / 2	104 x 104
2 x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			104 x 104
	Convolutional	256	3 x 3 / 2	52 x 52
8 x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			52 x 52
	Convolutional	512	3 x 3 / 2	26 x 26
8 x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			26 x 26
	Convolutional	1024	3 x 3 / 2	13 x 13
4 x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			13 x 13

13 x 13 x 1024

loss function:

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

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

Anchor → objectness score

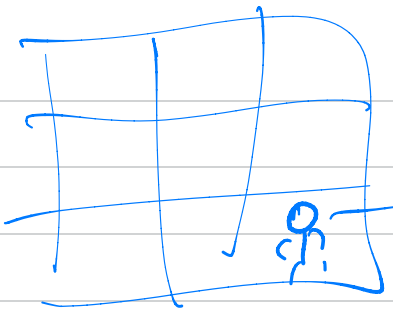
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2$$

8 -Vals

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

Grid → Class Prob Sc

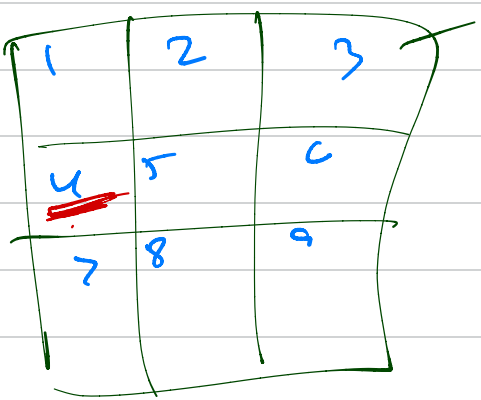
$$+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2$$



$(1-0.2)^2$ Total - 5
$(0-0.6)^2$
$+ (0-0.4)^2$

Objectness × Class Prob

IS NOT USED in Loss Function



3 × 3 Grid
5 - anchor box

→ 6th Grid cell
↓
3rd anchor box.

4H [0 0 0 0 0] → 0.1, 0.2, 0.3, 0.4, 0.1

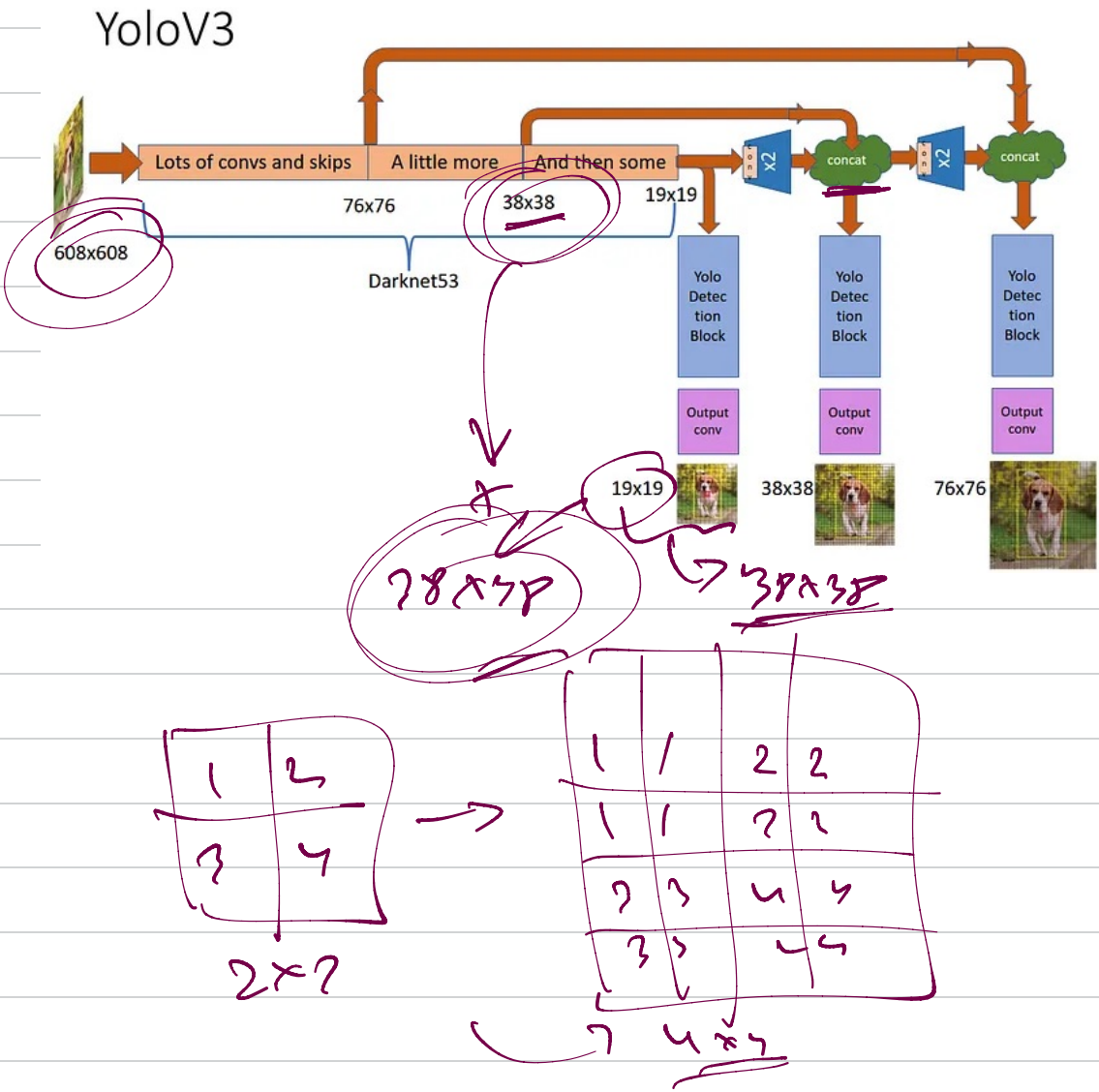
6H = [0 0 1 0 0]

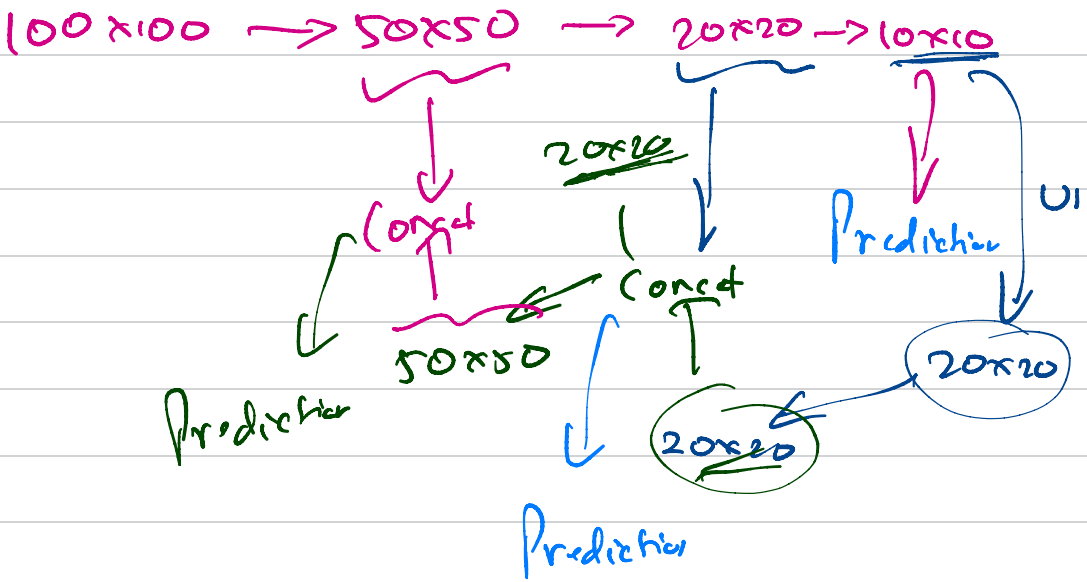
t_x
t_y
w_h
w_{-1}
} → Reg loss MSE

Objectness Score } → MSE (binary Cross entr

Class Prob } → Categorical Cross entropy

ARCH - YOLO V3


YoloV3
Lots of convs and skips
A little more
And then some
608x608
76x76
38x38
19x19
Darknet53
x2
concat
x2
concat
Yolo Detection Block
Yolo Detection Block
Yolo Detection Block
Output conv
Output conv
Output conv
19x19
38x38
76x76
28x28
38x38
2x2
4x4



20 x 20 x 100

+

20 x 20 x 512

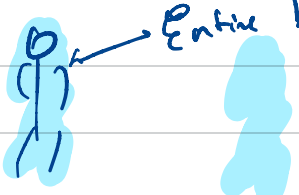
↓

20 x 20 x 612

RGBD → Depth

Semantic → Every single pixel is labelled
→ We don't have count of different instance of a class

→ Instance → Only pixels of chosen instance are labeled.
→ We have count of chosen instance.



Transpose Convolution

Input data → 224x224 → g'up

close to pixel

↓

Flatten

↓

$L_1 = [\ - - - - \ ]$

50176

Output → 224x224 (transpose conv)

↓

Flatten → L_2 [- - -]

50176

Accuracy = L_1 & L_2