

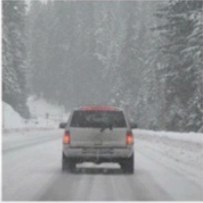
only one obj

tell me a bounding box for obj

multiple obj

What are localization and detection?

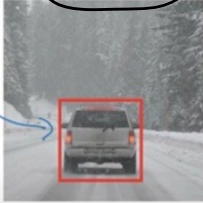
Image classification



"Car"

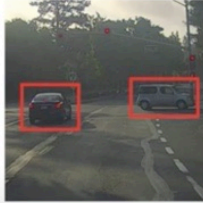
passport style img

Classification with localization



"Car"

Detection



C4W3L01 Object Localization

What is the difference between these problems?

Q: How will we be able to do localization?
 what input / what output

Input: $\{X, \text{is object?}, \text{midpoint + size box}, \text{class label}\}$

human annotator provides

output confidence of obj, bounding box, label of obj

Classification with Localization:


Defining the target label y

Need to output b_x, b_y, b_h, b_w , class label (1-4)

1 - pedestrian
 2 - car ←
 3 - motorcycle
 4 - background ←

$L(y, \hat{y}) = \begin{cases} (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \dots + (\hat{y}_n - y_n)^2 & \text{if } y_i = 1 \\ (\hat{y}_i - y_i)^2 & \text{if } y_i = 0 \end{cases}$

$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_w \\ b_h \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$ is there an object?

$x =$ 

$(x, y) \rightarrow \begin{bmatrix} 1 \\ b_x \\ b_y \\ b_w \\ b_h \\ 0 \\ \dots \\ 0 \end{bmatrix}$ ← "don't care"

Andrew Ng

What would the loss function be if you used cross-entropy loss for classification outputs and MSE for regression outputs?

Labels $y = (\underbrace{p_c}_w, b_x, b_y, b_w, b_h, \underbrace{c_1, c_2, c_3}_{\text{one hot encoding of class of object}})$

1 if obj pres
 0 if not

Output $\hat{y} = (\hat{p}_c, \hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h, \hat{c}_1, \hat{c}_2, \hat{c}_3)$

Binary CE regression MSE Classification CE

$$L(y, \hat{y}) = \begin{cases} -\log(1 - \hat{p}_c) & \text{if } p_c = 0 \\ -\log(\hat{p}_c) + \sum_i -1_{c_i=1} \log(\hat{c}_i) + \|b_x - \hat{b}_x\|^2 + \|b_y - \hat{b}_y\|^2 + \|b_w - \hat{b}_w\|^2 + \|b_h - \hat{b}_h\|^2 & \text{if } p_c = 1 \end{cases}$$

What is the theoretical underpinning of this loss?

Binary classification

$$L(\hat{y}, y) = \begin{cases} \log \hat{y} & \text{if } y=1 \\ -\log(1-\hat{y}) & \text{if } y=0 \end{cases}$$

$$= \mathbb{1}_{y=1} \log \hat{y} + \mathbb{1}_{y=0} \log(1-\hat{y})$$

Output

$$\hat{y} = (\hat{p}_c, \hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h, \hat{c}_1, \hat{c}_2, \hat{c}_3)$$

$\underbrace{\hspace{10em}}_{P(\text{object is there})} \quad \underbrace{\hspace{10em}}_{P(\text{object is at class } i \mid \text{object present})}$

Model? Given \mathcal{X} , there is a prob. dist of y

$$P(\text{class } i \mid \mathcal{X}) = \hat{p}_c \hat{c}_i = P(\text{class } i \mid \text{object}, \mathcal{X}) \times P(\text{object} \mid \mathcal{X})$$

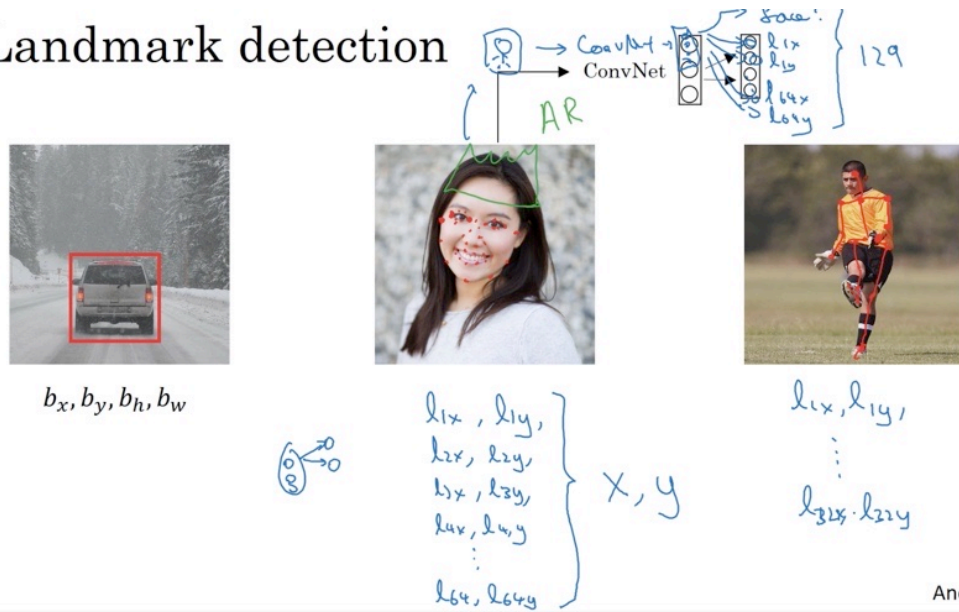
$$\begin{matrix} b_x \mid \mathcal{X} = \hat{b}_x + \epsilon \mathcal{N}(0,1) \\ \vdots \\ \vdots \end{matrix}$$

Max Likelihood?

$$L(\theta) = \prod_i \left[\hat{p}_c \hat{c}_1^{\mathbb{1}_{c_1=1}} \hat{c}_2^{\mathbb{1}_{c_2=1}} \hat{c}_3^{\mathbb{1}_{c_3=1}} e^{-\|b_x - \hat{b}_x\|^2 - \|b_y - \hat{b}_y\|^2} \right]^{p_c} (1 - \hat{p}_c)^{1-p_c}$$

likelihood params

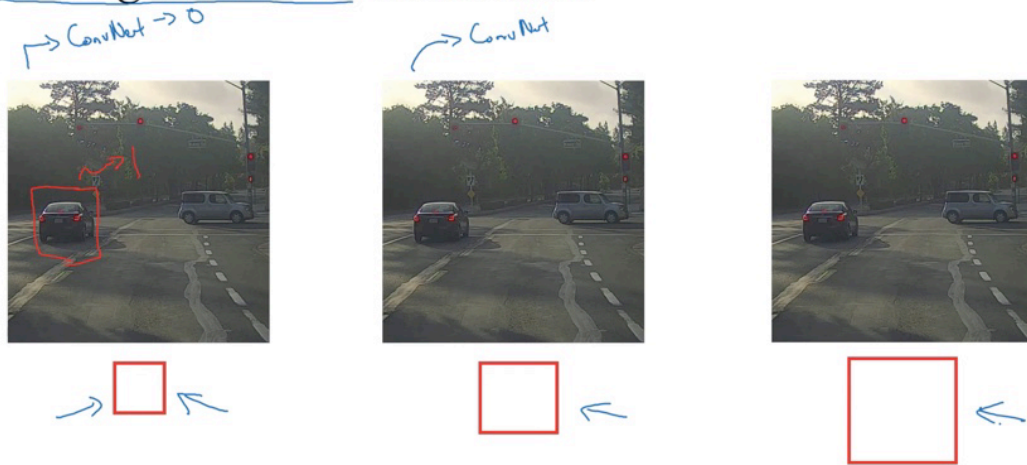
Landmark detection



What are some concerns about doing landmark detection as described?

- Noise in location a person would tag
- May be ambiguity in what a landmark means

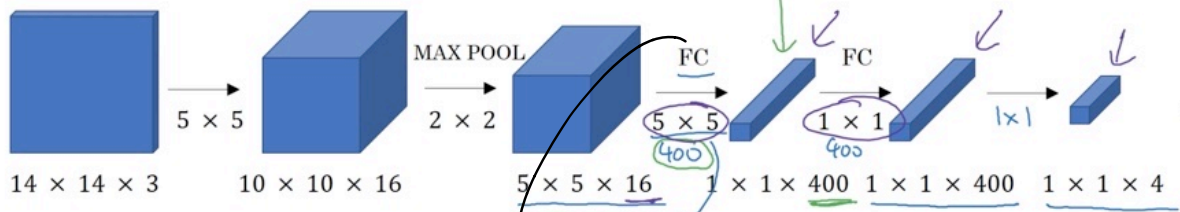
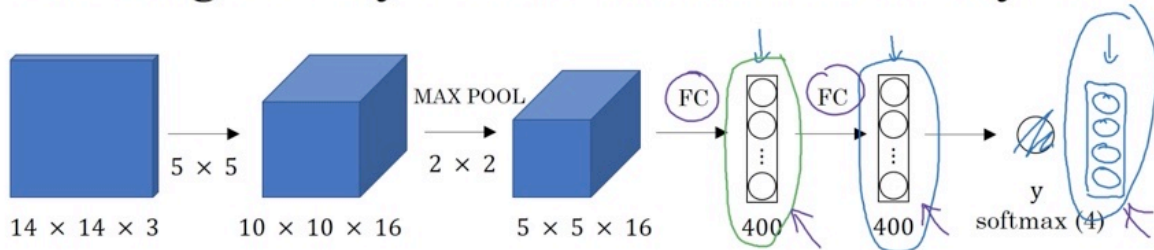
Sliding windows detection



Andrew Ng

How do you pass larger image fragments into the same classifier?

Turning FC layer into convolutional layers



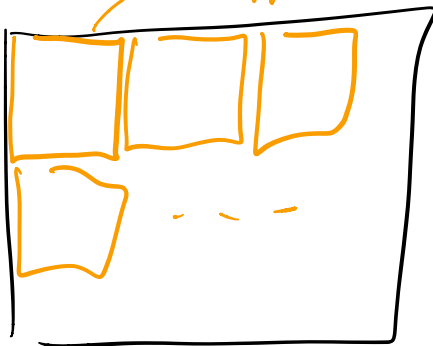
Will the output be exactly equal?
YES

5x5 conv w/ 400 output channels

Andrew Ng

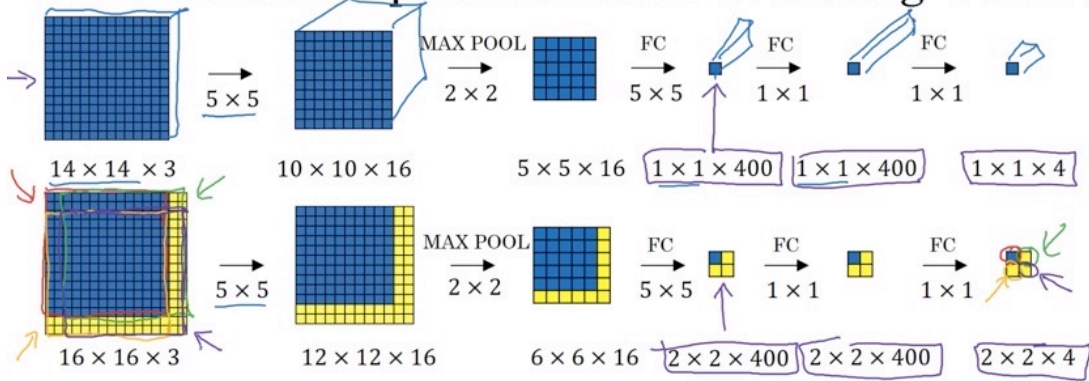
By changing these fully connected layers into convolutional layers, what does that now allow us to do?

Can now apply net to larger images
net is applied to this



Each output dim is the same as if original input image was cropped

Convolution implementation of sliding windows

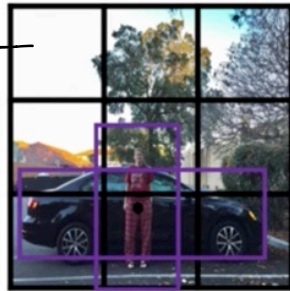


[Sermanet et al., 2014, OverFeat: Integrated recognition, localization and detection using convolutional networks]

Andrew Ng

Overlapping objects:

Each grid cells, predict if there is an obj & if so, where

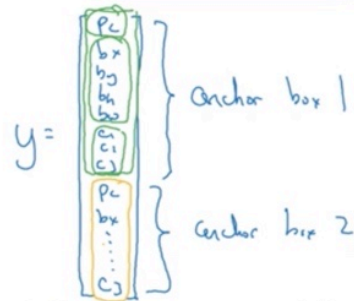


$$y = \begin{bmatrix} p_c \\ b_x \\ b_y \\ b_h \\ b_w \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

Anchor box 1:



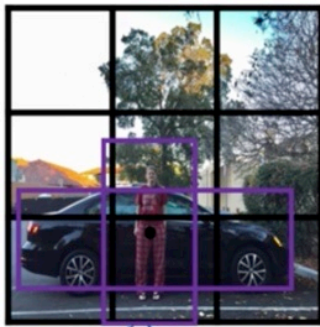
Anchor box 2:



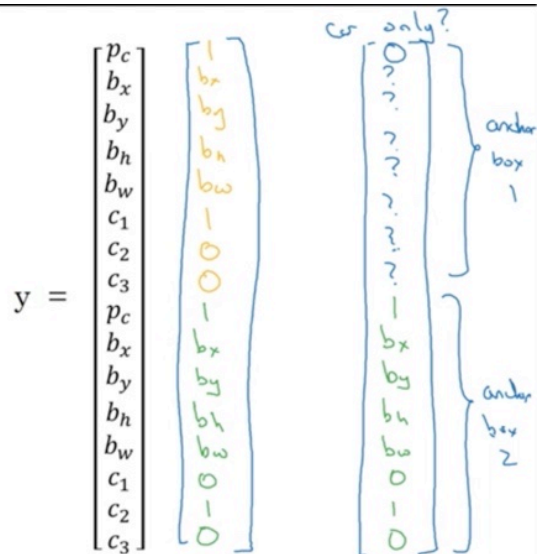
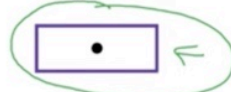
[Redmon et al., 2015, You Only Look Once: Unified real-time object detection]

Andrew Ng

Anchor box example



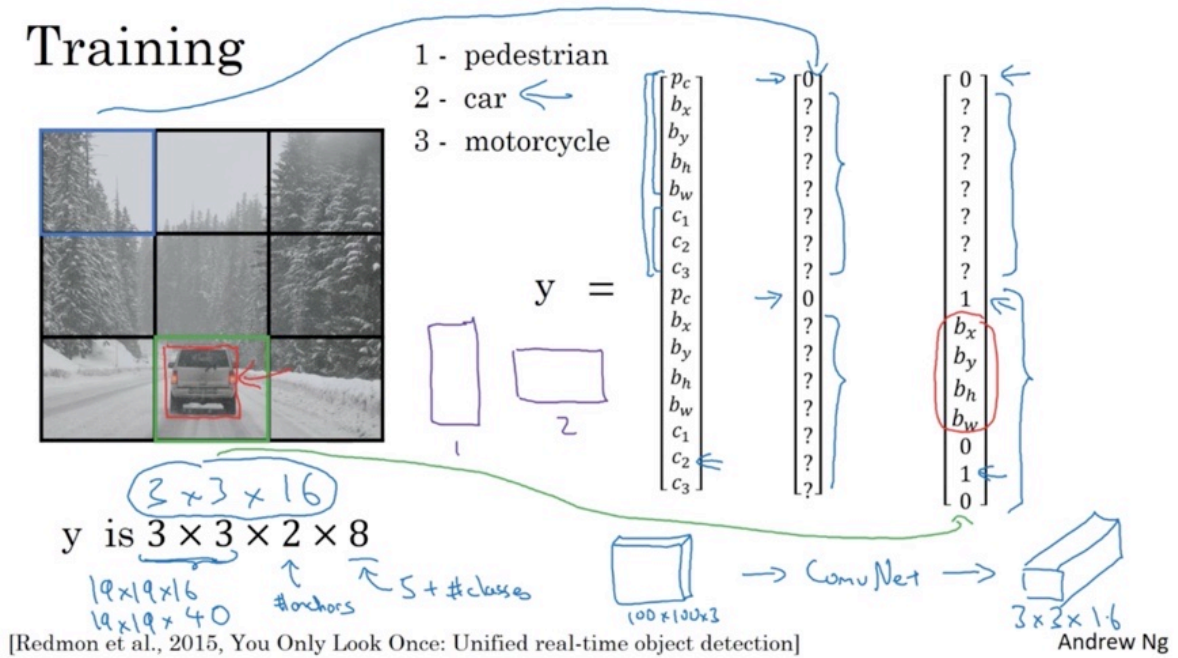
Anchor box 1: Anchor box 2:



Andrew Ng

YOLO Training

Training



Making predictions

