AI VIETNAM
All-in-One Course

Extra Class

YOLO – V1

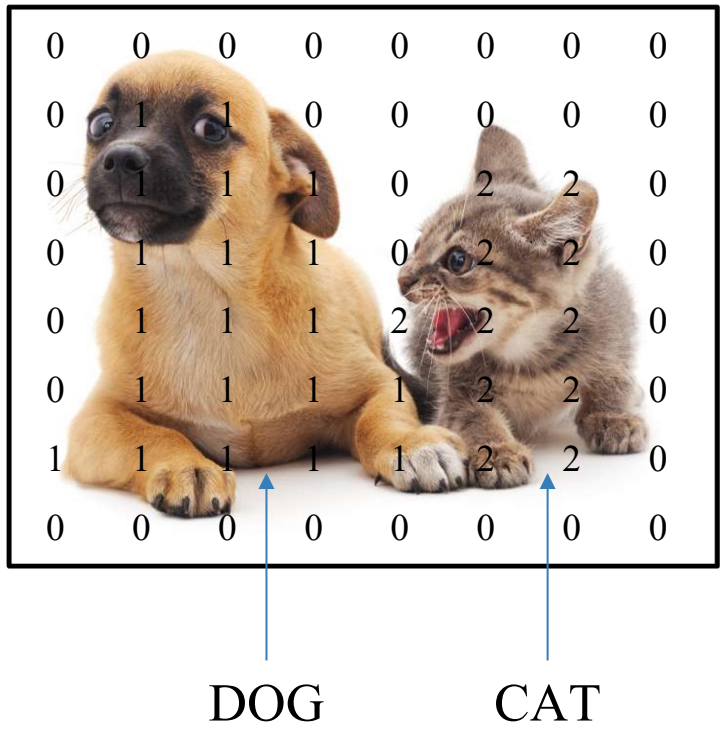
Nguyen Quoc Thai

CONTENT

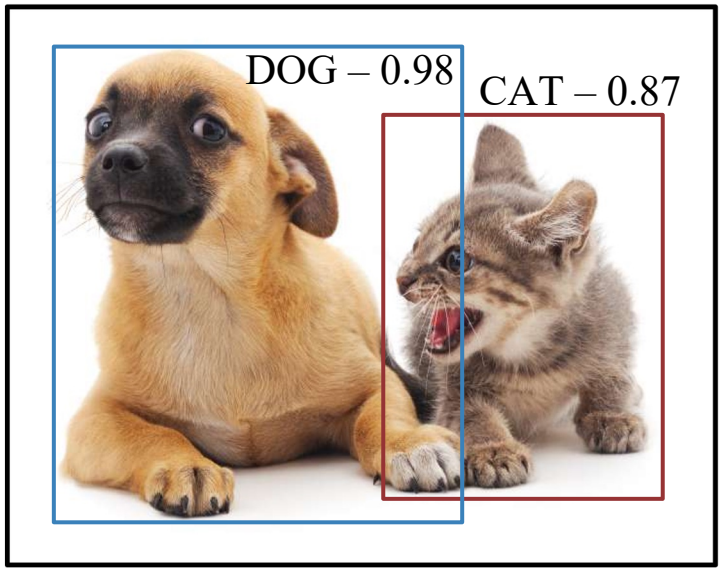
1 – Object Detection

Object Detection

Image Segmentation



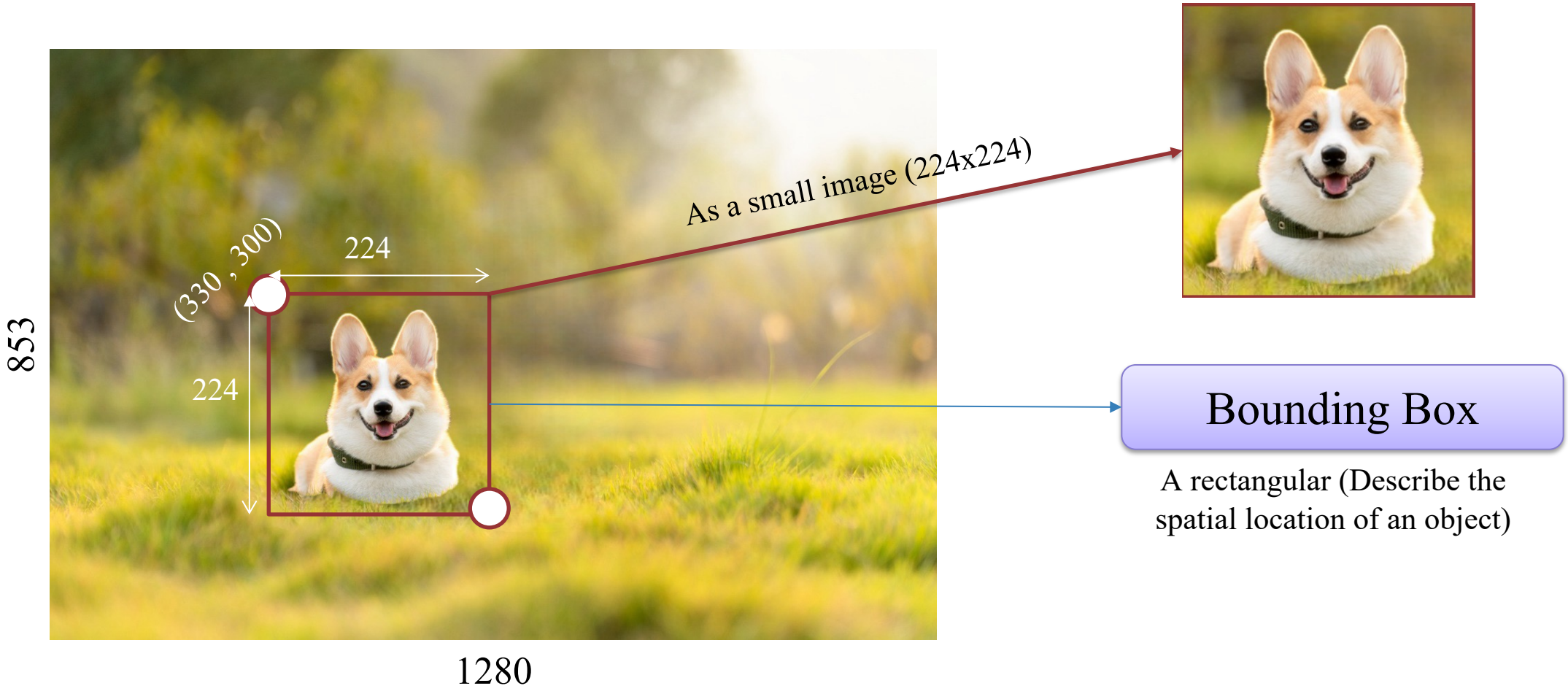
Object Detection



Assign labels, bounding boxes to objects in the image

1 – Object Detection

Naïve Object Detection



Bounding Box

A rectangular (Describe the spatial location of an object)

1 – Object Detection

Naïve Object Detection

AI VIET NAM
@aivietnam.edu.vn

2 – YOLO V1

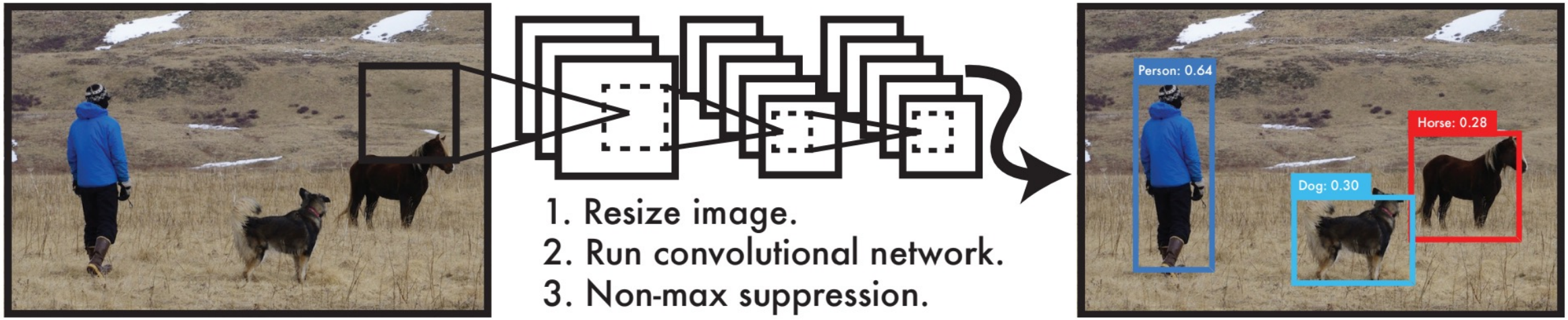
YOLO – V1

You Only Look Once
Unified, Real-Time Object Detection

(1) A single neural network for localization and for classification

(2) Need to inference only once

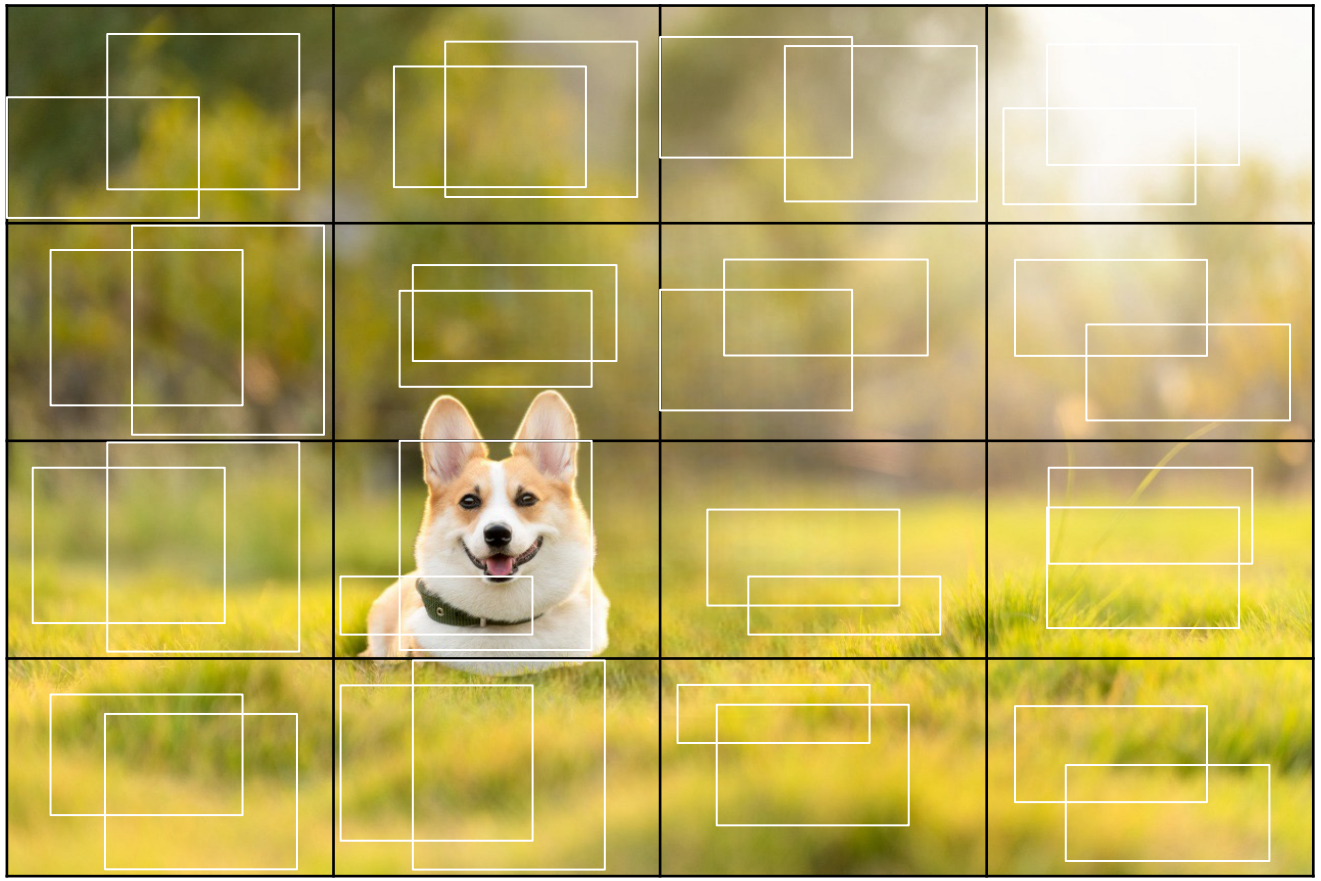
(3) Looks at the entire image each time leading to less false positives



2 – YOLO V1

Unified Detection

- Split into a S x S grid
- For each grid square, generate B bounding boxes
- For each bounding box, 5 predictions: (x, y, w, h, confidence)



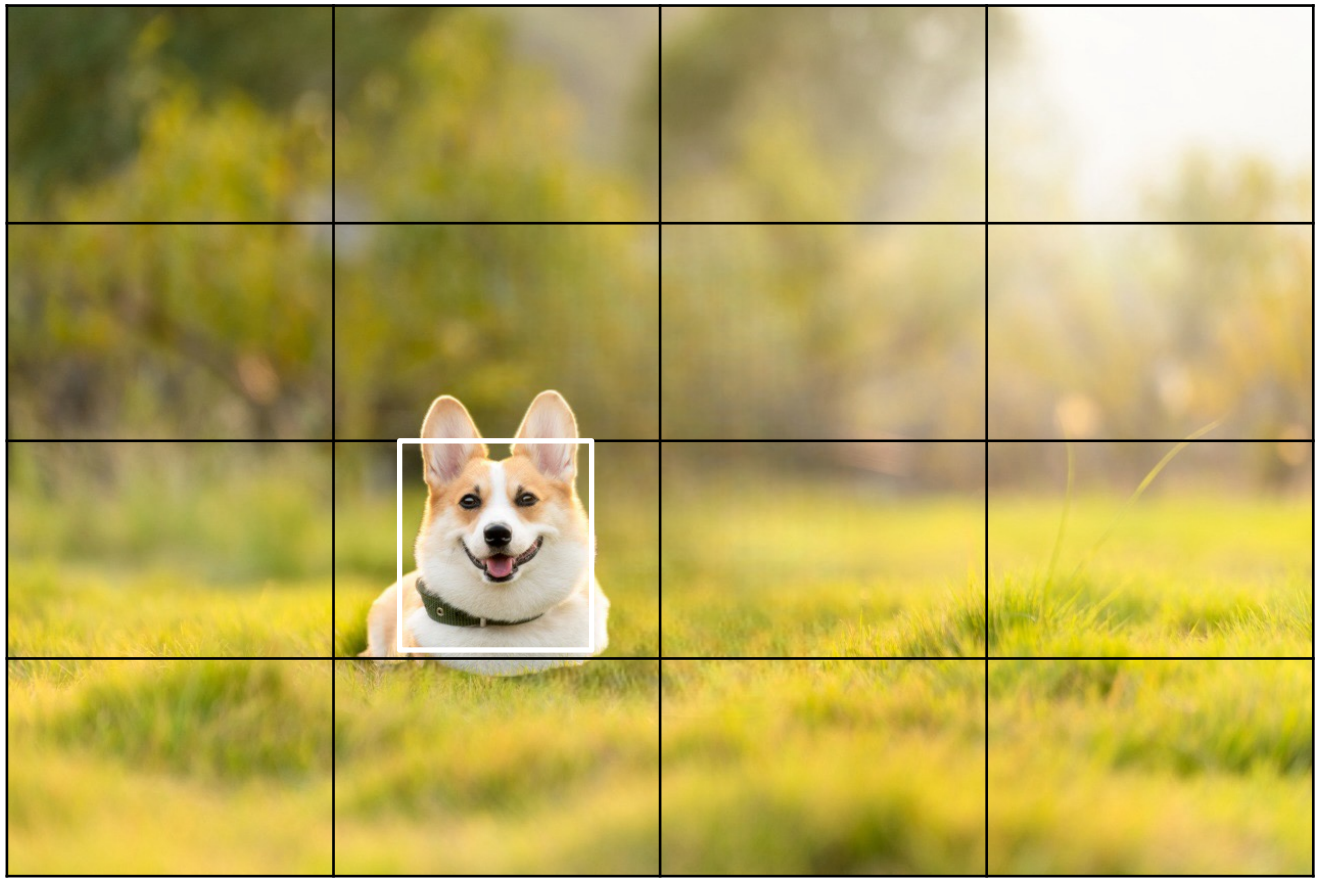
2 – YOLO V1

Unified Detection

- Split into a S x S grid
- For each grid square, generate B bounding boxes
- For each bounding box, 5 predictions: (x, y, w, h, confidence)

- (x, y): the center of the box relative to the bounds of the grid
- $h = \frac{h_{box}}{H_{image}}$
- $w = \frac{w_{box}}{W_{image}}$
- confidence score

$$\Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}}$$



2 – YOLO V1

Unified Detection – Demo

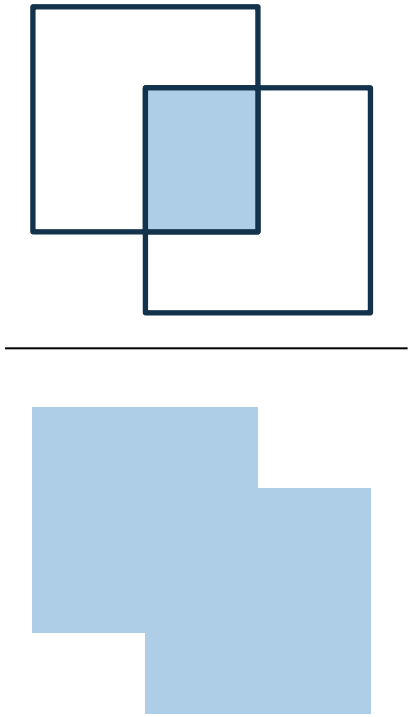
- Split into a S x S grid
- For each grid square, generate B bounding boxes
- For each bounding box, 5 predictions:

(x, y, w, h, confidence)

- (x, y): the center of the box relative to the bounds of the grid
- $h = \frac{h_{box}}{H_{image}}$
- $w = \frac{w_{box}}{W_{image}}$
- confidence score

$$\Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}}$$

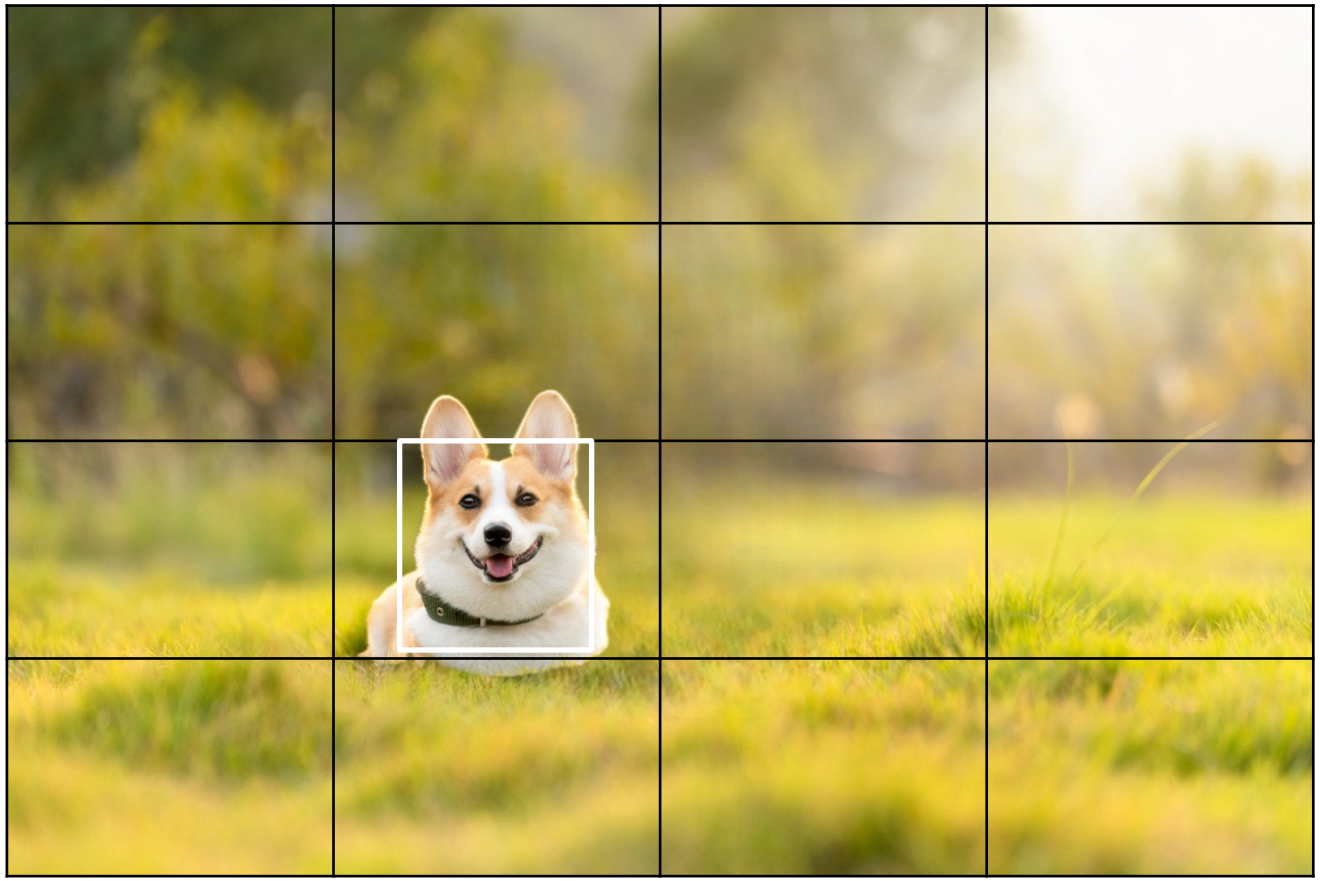
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} =$$



2 – YOLO V1

Unified Detection

- Split into a S x S grid
- For each grid square, generate B bounding boxes
- For each bounding box, 5 predictions: (x, y, w, h, confidence)
- C: conditional class probabilities: $\Pr(\text{Class}_i|\text{Object})$

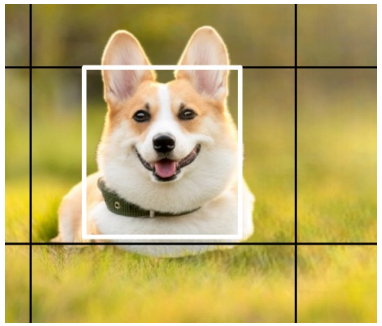


2 – YOLO V1

Unified Detection

- Split into a S x S grid
- For each grid square, generate B bounding boxes
- For each bounding box, 5 predictions: (x, y, w, h, confidence)
- C: conditional class probabilities: $Pr(Class_i|Object)$

YOLO is a regression algorithm

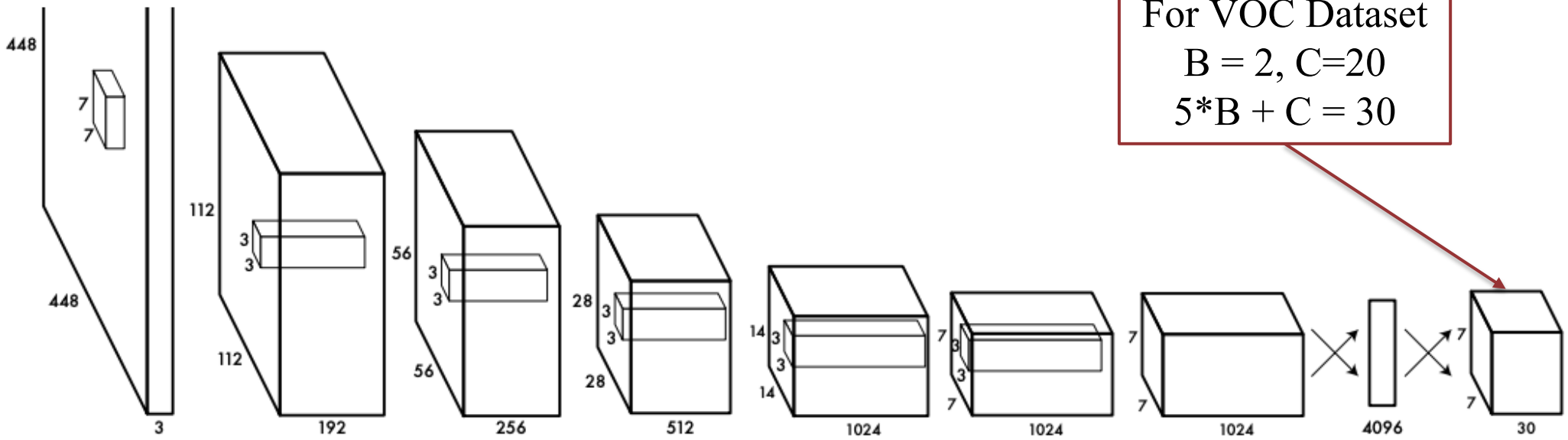
X		Y	
	**W * H * 3** Width Height RGB Values		**S * S * (B*5 + C)** (B*5 + C) predictions + class predicted distribution for a gird block

AI VIET NAM
@aivietnam.edu.vn

2 – YOLO V1

Network Design – Demo

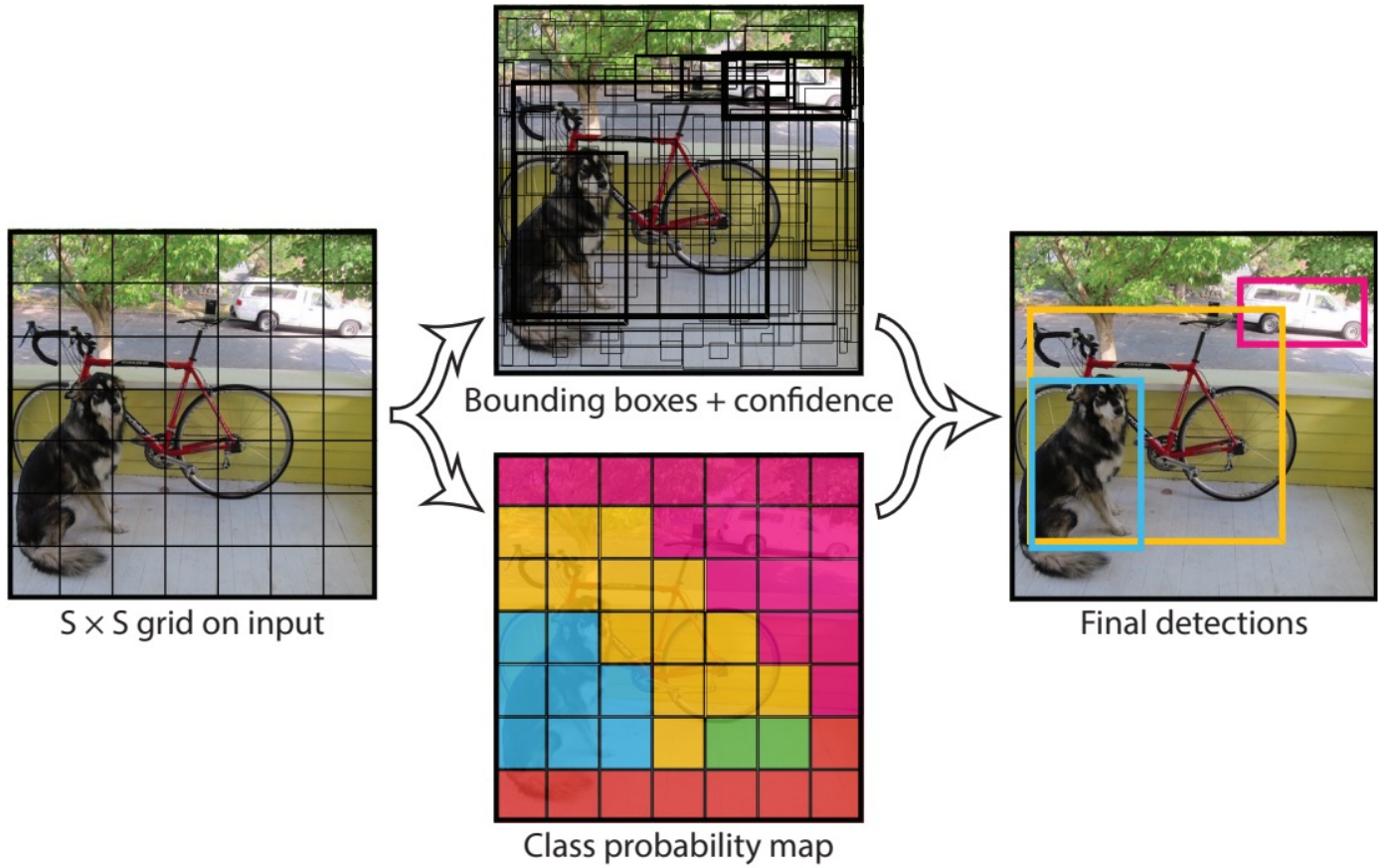
For VOC Dataset
B = 2, C=20
5*B + C = 30



2 – YOLO V1

Inference

- Use a threshold to filter out bouding boxes with low P(Object)
- Multiply the conditional class probabilities and the individual box confidence predictions

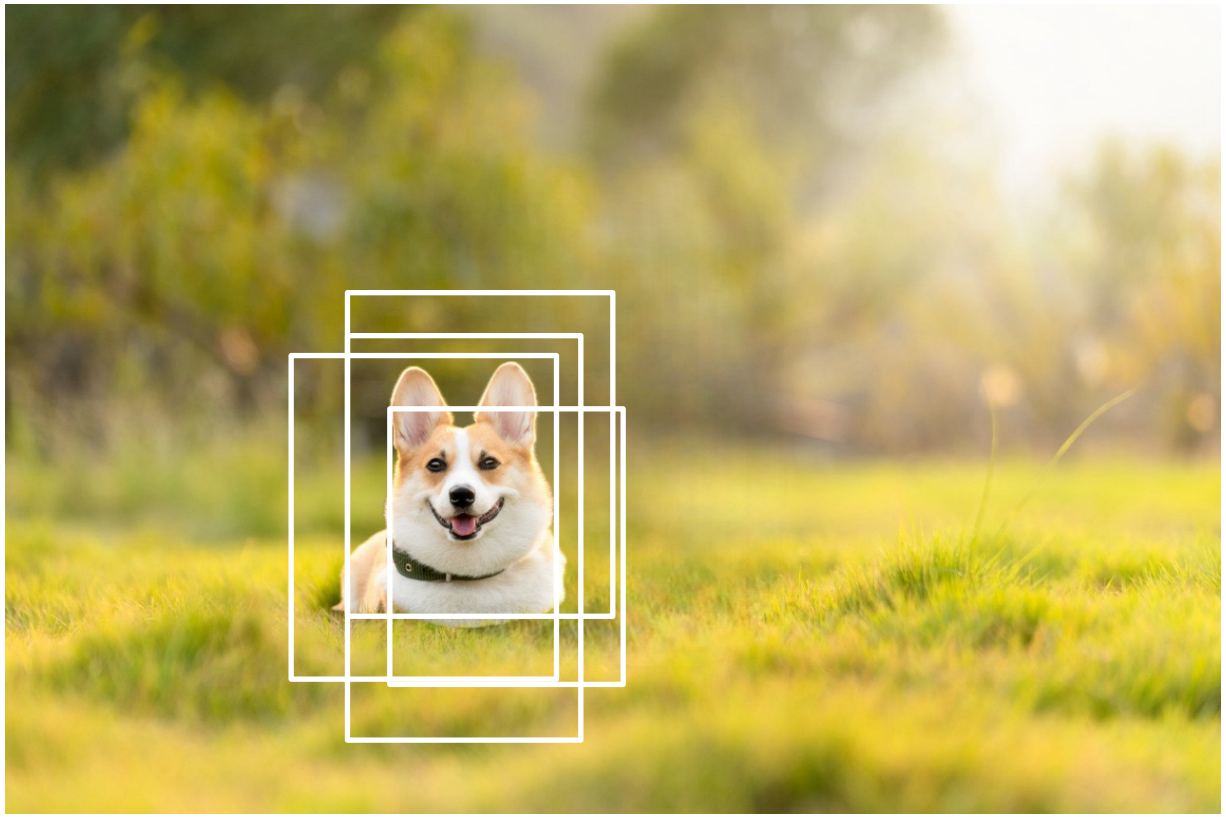


$$\mathbf{Pr(Class_i|Object) * Pr(Object) * IOU^{truth}_{pred} = Pr(Class_i) * IOU^{truth}_{pred}}$$

2 – YOLO V1

Non-maximal Suppresion

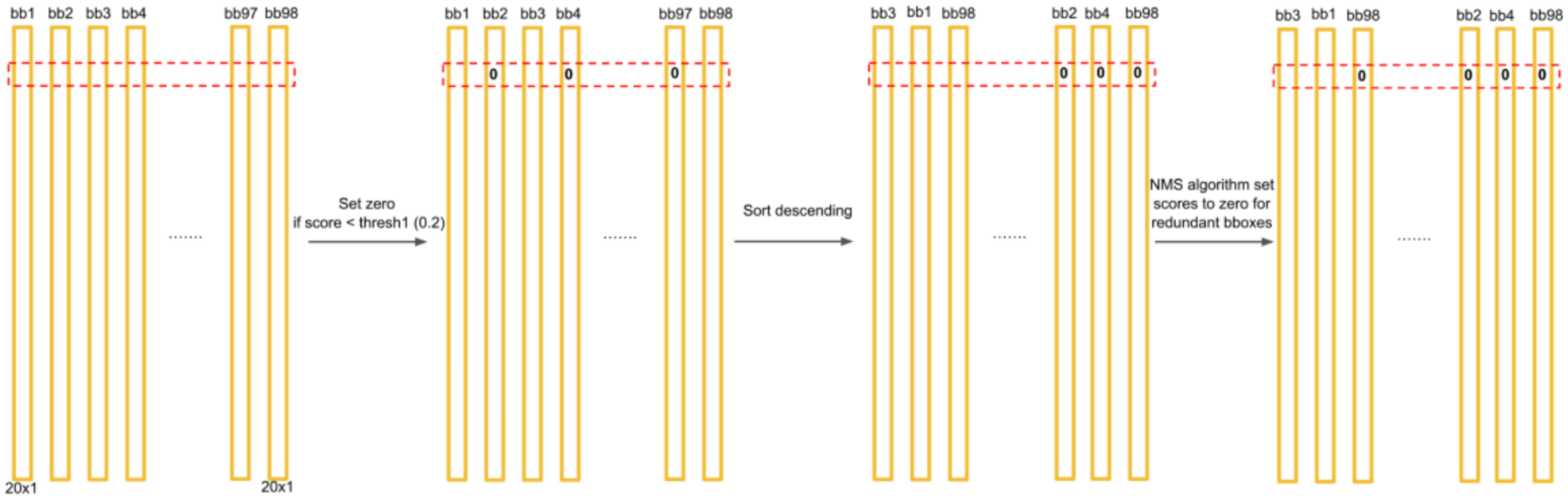
- Discard bounding box with high overlap (keep the bounding box with highest confidence)



2 – YOLO V1

Non-maximal Suppresion – Demo

- Discard bounding box with high overlap (keep the bounding box with highest confidence)
- $\Pr(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}}$

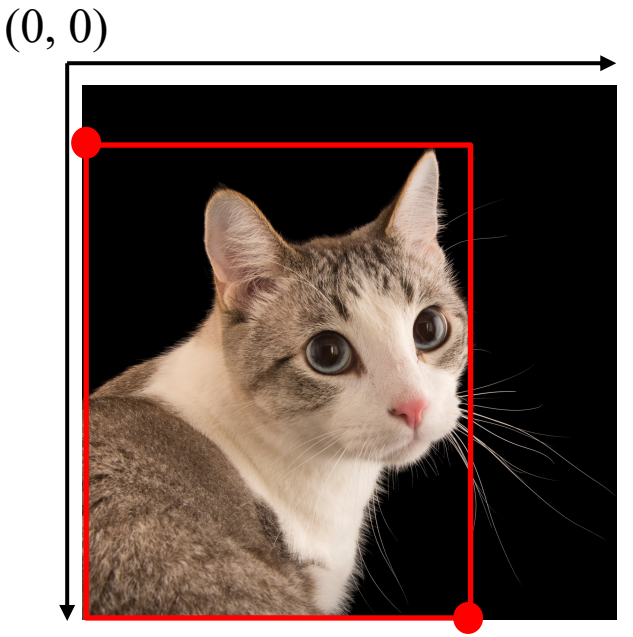


AI VIET NAM
@aivietnam.edu.vn

3 – VOC Dataset

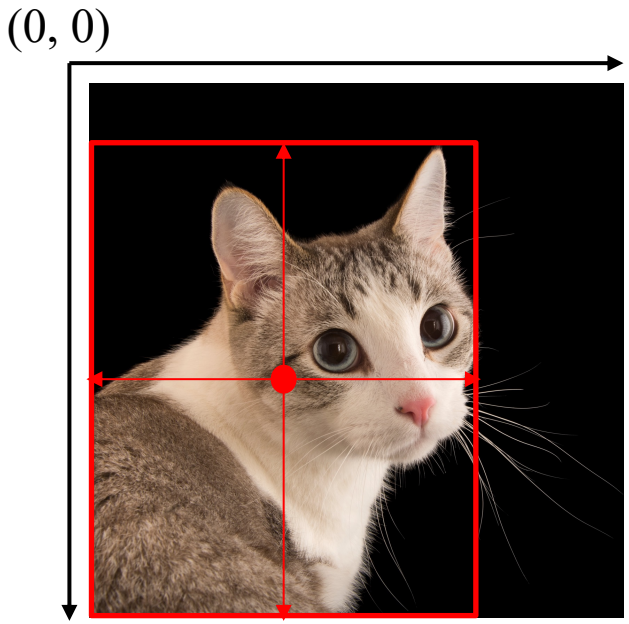
Object Detection Dataset Format

VOC



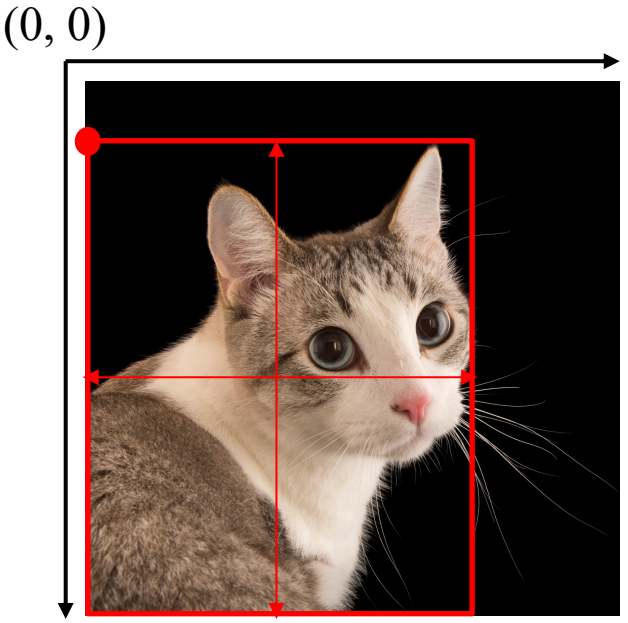
$(x_{min}, y_{min}, x_{max}, y_{max})$

YOLO



$(x_{center}, y_{center}, width, height)$

COCO



$(x_{min}, y_{min}, width, height)$

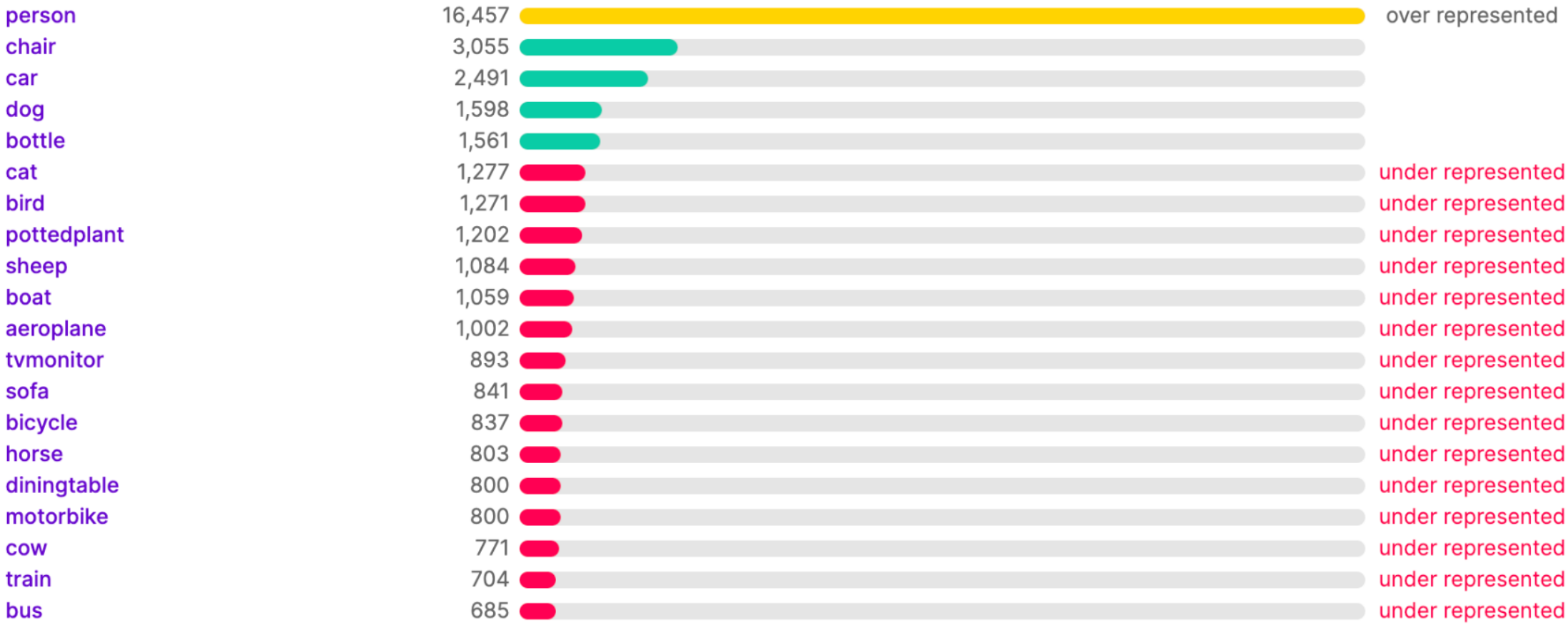
3 – VOC Dataset

Pascal VOC Detection Dataset

- ***Person**:* person
- ***Animal**:* bird, cat, cow, dog, horse, sheep
- ***Vehicle**:* aeroplane, bicycle, boat, bus, car, motorbike, train
- ***Indoor**:* bottle, chair, dining table, potted plant, sofa, tv/monitor

3 – VOC Dataset

Pascal VOC Detection Dataset



3 – VOC Dataset

Pascal VOC Detection Dataset

VOCDETECTION

CLASS torchvision.datasets.VOCDetection(*root: str, year: str = '2012', image_set: str = 'train', download: bool = False, transform: Optional[Callable] = None, target_transform: Optional[Callable] = None, transforms: Optional[Callable] = None*) [SOURCE]

Pascal VOC Detection Dataset.

Parameters:

- **root** (*string*) – Root directory of the VOC Dataset.
- **year** (*string, optional*) – The dataset year, supports years `"2007"` to `"2012"`.
- **image_set** (*string, optional*) – Select the image_set to use, `"train"`, `"trainval"` or `"val"`. If `year=="2007"`, can also be `"test"`.
- **download** (*bool, optional*) – If true, downloads the dataset from the internet and puts it in root directory. If dataset is already downloaded, it is not downloaded again. (default: alphabetic indexing of VOC's 20 classes).
- **transform** (*callable, optional*) – A function/transform that takes in an PIL image and returns a transformed version. E.g, `transforms.RandomCrop`
- **target_transform** (*callable, required*) – A function/transform that takes in the target and transforms it.
- **transforms** (*callable, optional*) – A function/transform that takes input sample and its target as entry and returns a transformed version.

3 – VOC Dataset

Pascal VOC Detection Dataset

Image with 1 bounding box

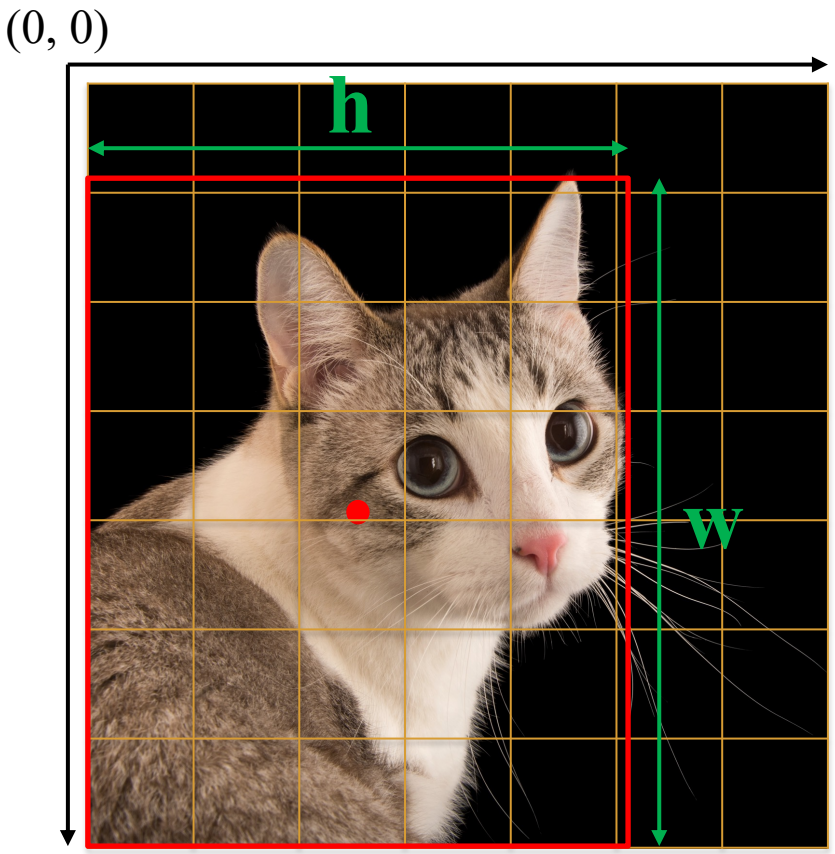
```
<annotation>
        <folder>VOC2012</folder>
        <filename>2008_000616.jpg</filename>
        <source>
                <database>The VOC2008 Database</database>
                <annotation>PASCAL VOC2008</annotation>
                <image>flickr</image>
        </source>
        <size>
                <width>231</width>
                <height>256</height>
                <depth>3</depth>
        </size>
        <segmented>0</segmented>
        <object>
                <name>person</name>
                <bndbox>
                        <xmin>51</xmin>
                        <ymin>62</ymin>
                        <xmax>209</xmax>
                        <ymax>256</ymax>
                </bndbox>
        </object>
</annotation>
```

Image with 2 bounding boxes

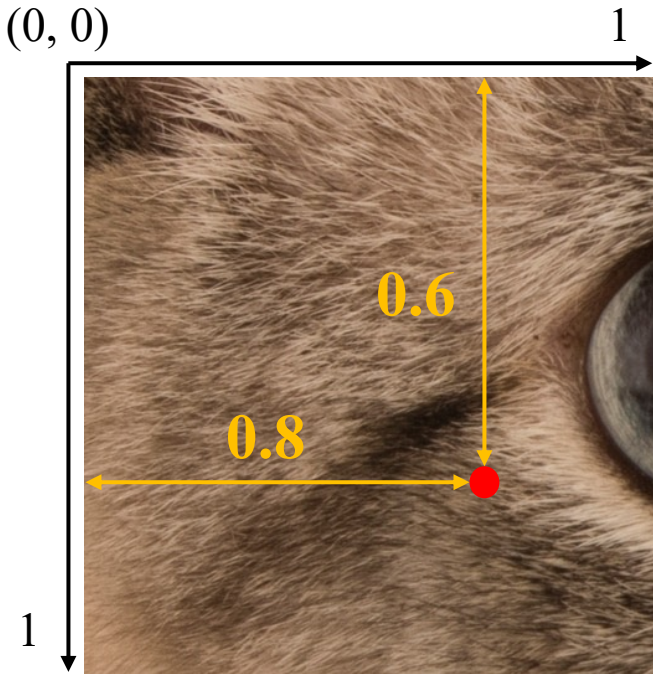
```
<annotation>
        <folder>VOC2012</folder>
        <filename>2008_000200.jpg</filename>
        <source>
                <database>The VOC2008 Database</database>
                <annotation>PASCAL VOC2008</annotation>
                <image>flickr</image>
        </source>
        <size>
                <width>500</width>
                <height>375</height>
                <depth>3</depth>
        </size>
        <segmented>0</segmented>
        <object>
                <name>person</name>
                <bndbox>
                        <xmin>119</xmin>
                        <ymin>76</ymin>
                        <xmax>184</xmax>
                        <ymax>311</ymax>
                </bndbox>
        </object>
        <object>
                <name>person</name>
                <bndbox>
                        <xmin>266</xmin>
                        <ymin>43</ymin>
                        <xmax>338</xmax>
                        <ymax>323</ymax>
                </bndbox>
        </object>
</annotation>
```

3 – VOC Dataset

Rescale Bounding Box



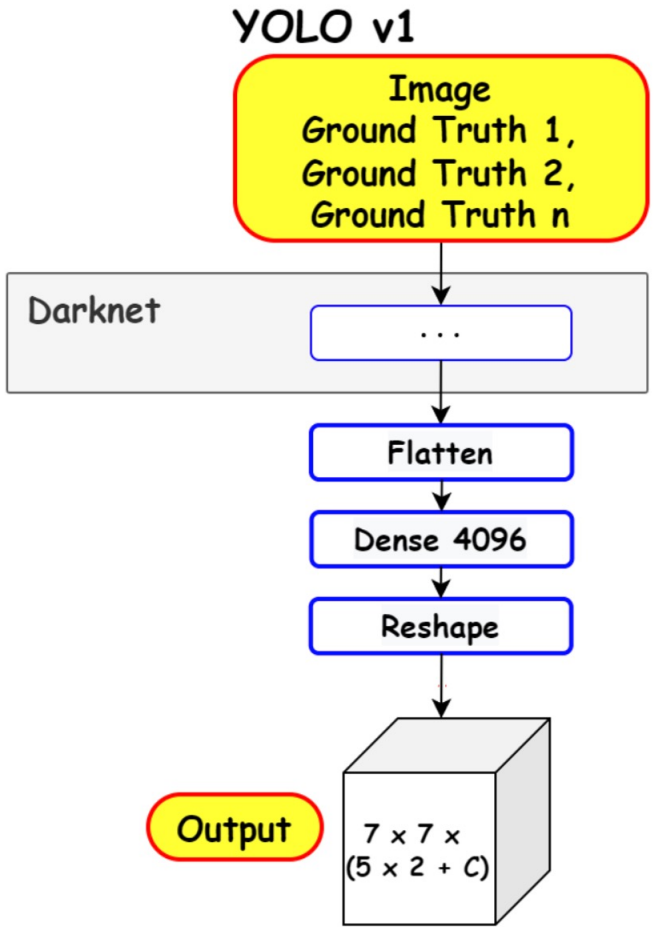
rescale

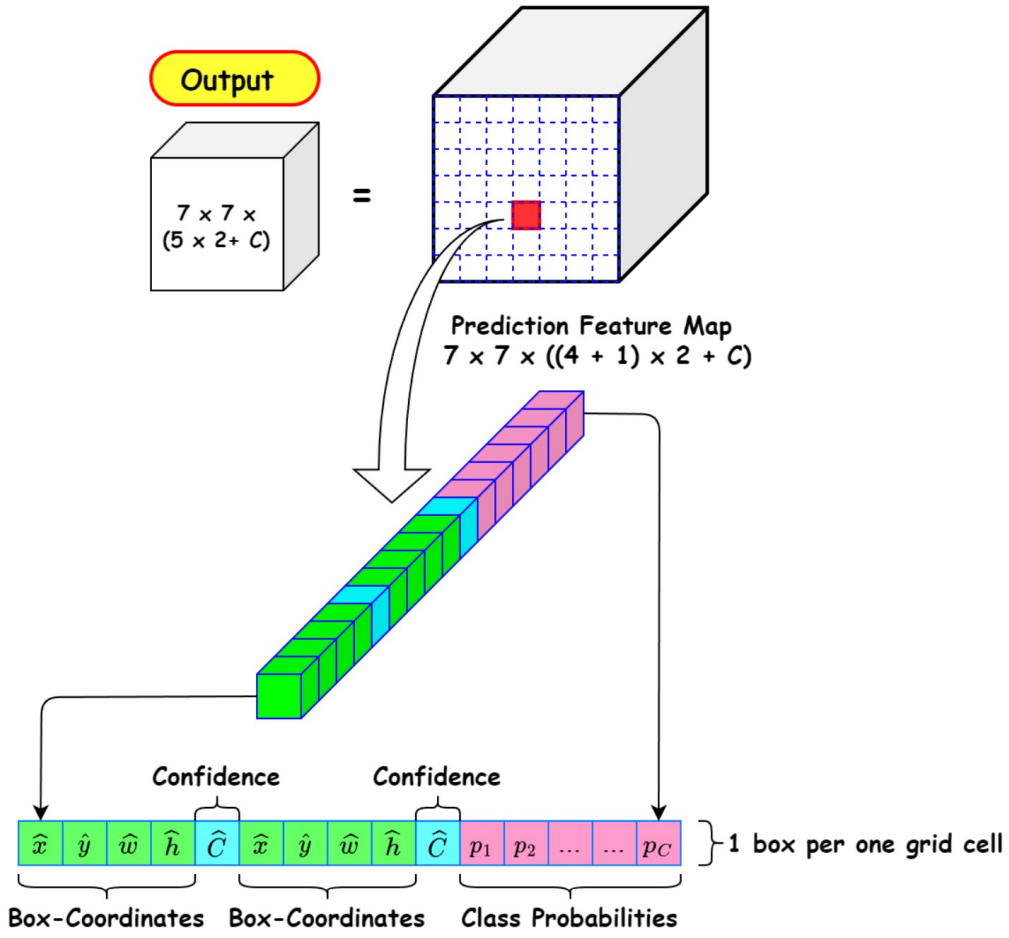


x	y	w	h
0.8	0.6	5.1	6.2

3 – VOC Dataset

Output Format

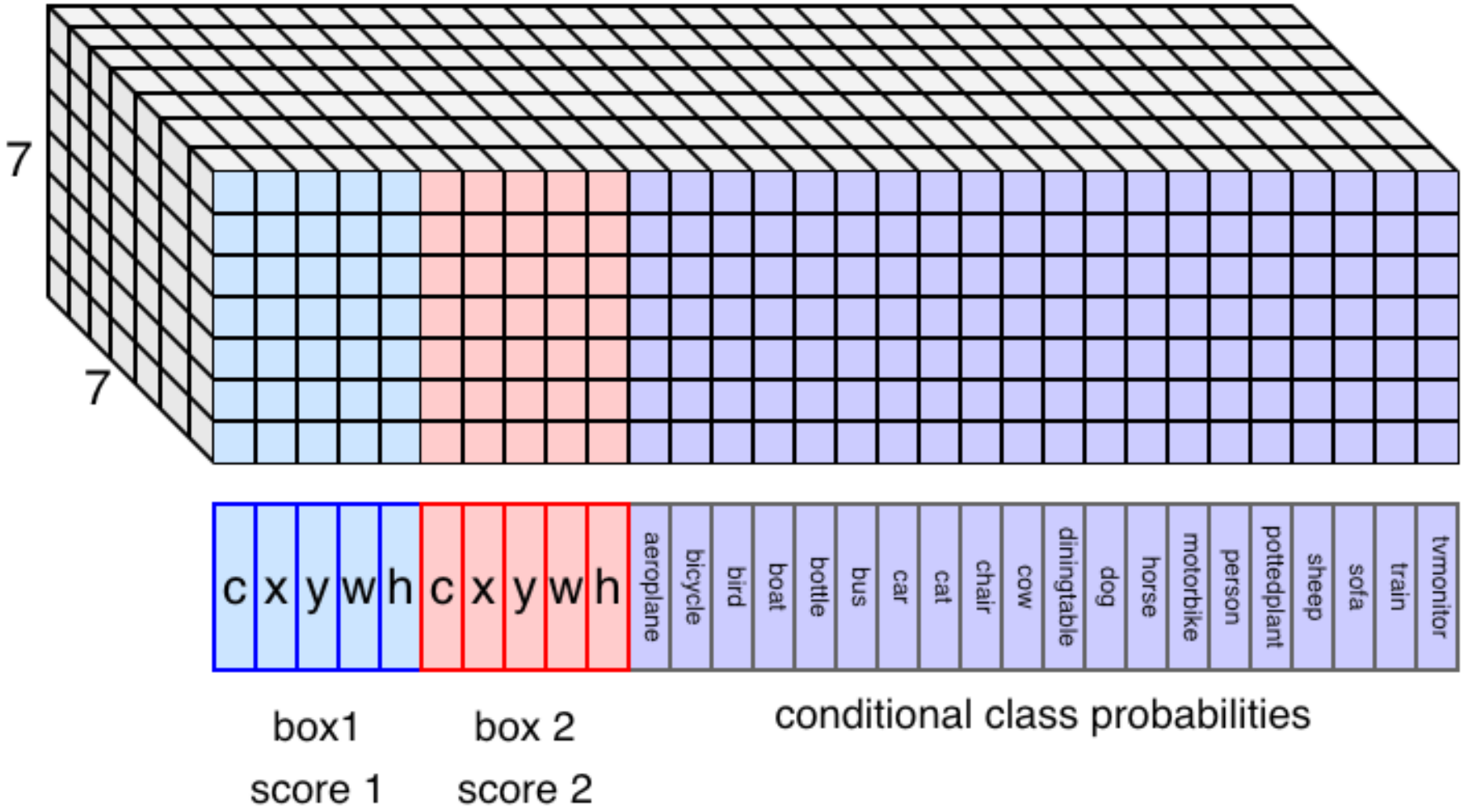




AI VIET NAM
@aivietnam.edu.vn

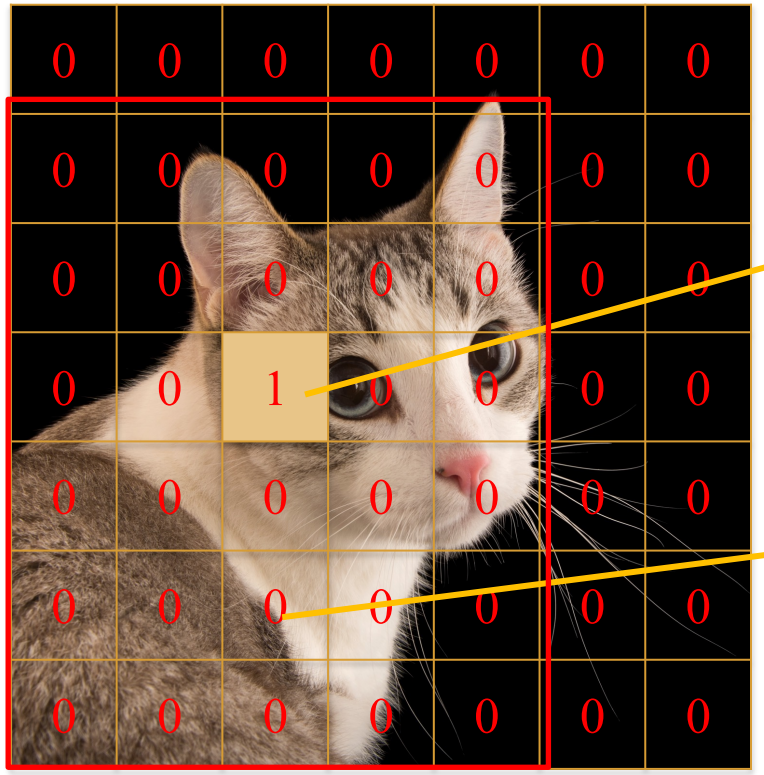
3 – VOC Dataset

Output Shape

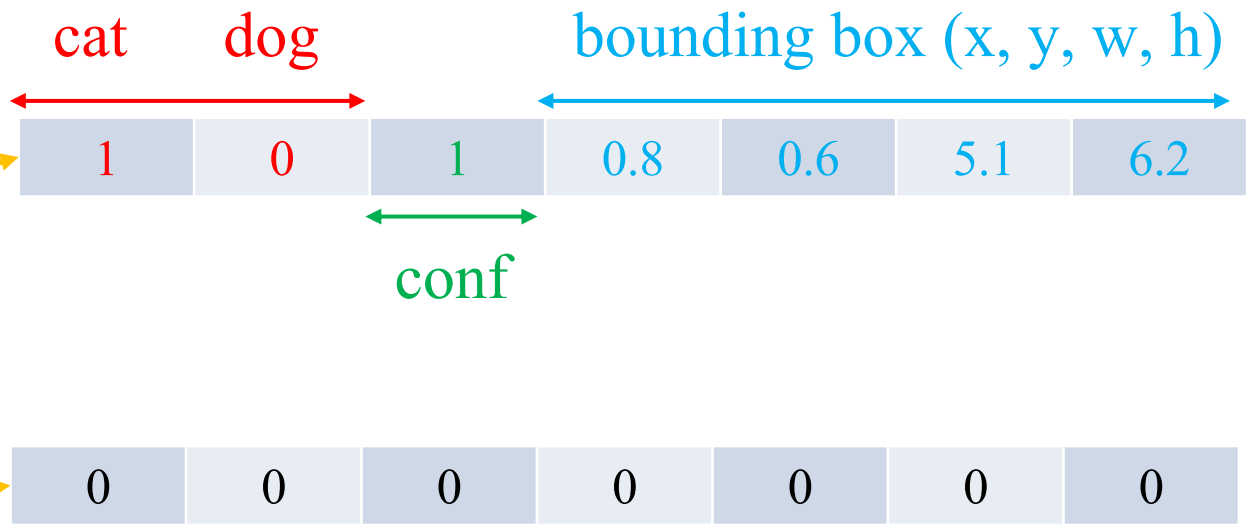


3 – VOC Dataset

Target for Training



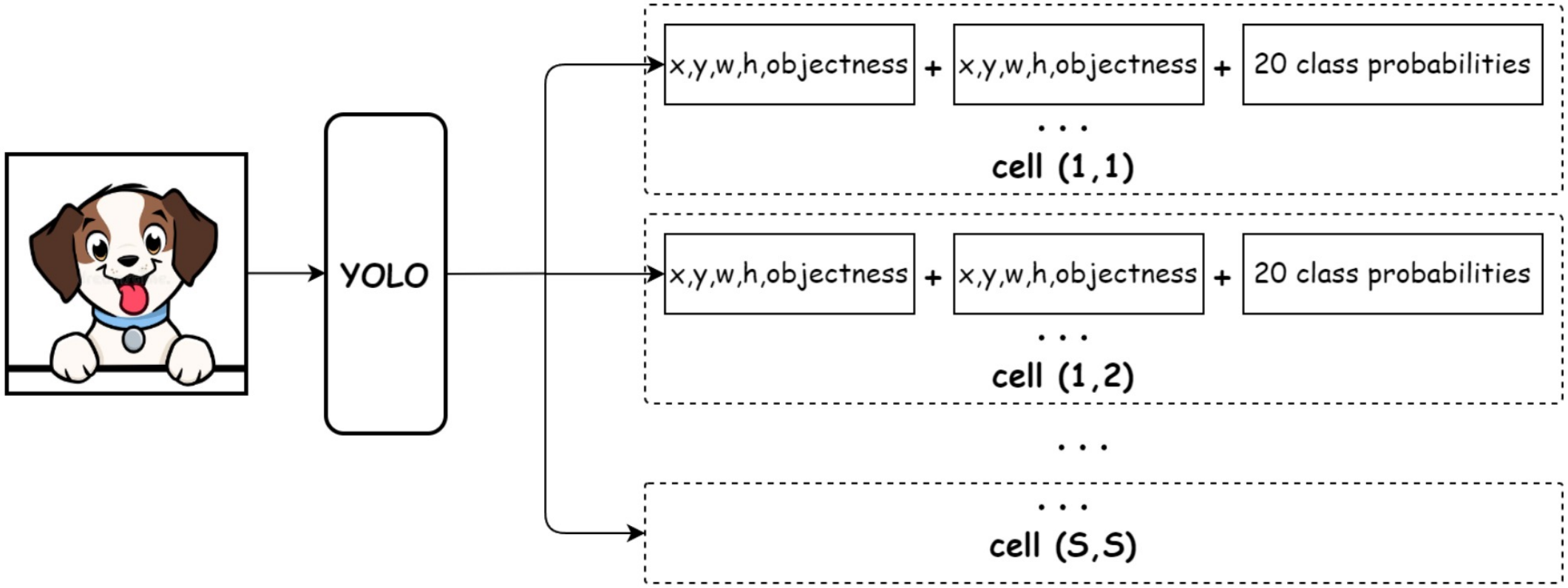
For example, we have 2 classes: *cat* and *dog*.



For each object, **only 1 grid cell** is responsible for the prediction.

3 – VOC Dataset

Output Example



3 – VOC Dataset

Ground Truth Example

	cat	dog	conf	bounding box (x, y, w, h)			
cell 1	1	0	1	0.8	0.6	5.1	6.2
cell 2	0	0	0	0	0	0	0

3 – VOC Dataset

Localization loss

The localization loss measures the errors in the predicted boundary box locations and sizes. We only count the box responsible for detecting the object.

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

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

where

$1_{ij}^{obj} = 1$ if the j th boundary box in cell i is responsible for detecting the object, otherwise 0.

λ_{coord} increase the weight for the loss in the boundary box coordinates.

$\lambda_{coord} = 5$

	cat	dog	conf	bounding box (x, y, w, h)			
cell 1	1	0	1	0.8	0.6	5.1	6.2
cell 2	0	0	0	0	0	0	0
cell 1	0.8	0.2	0.8	0.7	0.65	3.1	7.2
cell 2	0.45	0.55	0.2	0.35	0.2	0.5	1.2

$$\mathcal{L}_{loc} = \lambda_{coord}[(0.8 - 0.7)^2 + (0.6 - 0.65)^2 + (\sqrt{5.1} - \sqrt{3.1})^2 + (\sqrt{6.2} - \sqrt{7.2})^2]$$

3 – VOC Dataset

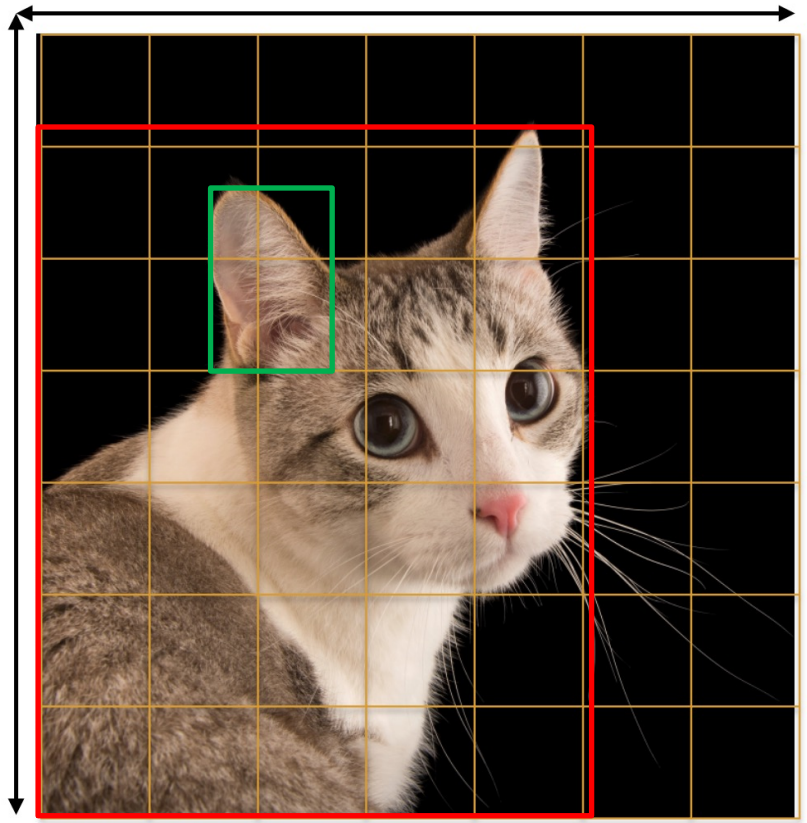
Why we use Square Root for width and height in localization loss

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

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

width_gt	width_pred	Error normal	Square Root Error
5.5	5.0	0.5	0.11
1.0	0.5	0.5	0.29

The difference between a small box is much more significant than a large box.



3 – VOC Dataset

Confidence loss

If an object is detected in the box, the confidence loss (measuring the objectness of the box) is:

$$\mathcal{L}_{\text{obj}} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} 1_{ij}^{\text{obj}}(C_{ij} - \hat{C}_{ij})^2$$

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

where

$\hat{C}_i$ is the box confidence score of the box j in cell i.

$1_{ij}^{obj} = 1$ if the j th boundary box in cell i is responsible for detecting the object, otherwise 0.

$\lambda_{noobj} = 0.5$

	cat	dog	conf	bounding box			
cell 1	1	0	1	0.8	0.6	5.1	6.2
cell 2	0	0	0	0	0	0	0
cell 1	0.8	0.2	0.8	0.7	0.65	3.1	7.2
cell 2	0.45	0.55	0.2	0.35	0.2	0.5	1.2

$$\mathcal{L}_{obj} = (1 - 0.8)^2 + \lambda_{noobj}(0.0 - 0.2)^2$$

3 – VOC Dataset

Classification loss

If an object is detected, the classification loss at each cell is the squared error of the class conditional probabilities for each class:

$$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{S^2} 1_i^{\text{obj}} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2$$

where

$\mathbb{1}_i^{obj} = 1$ if an object appears in cell i, otherwise 0.

$\hat{p}_i(c)$ denotes the conditional class probability for class c in cell i.

	cat	dog	conf	bounding box			
cell 1	1	0	1	0.8	0.6	5.1	6.2
cell 2	0	0	0	0	0	0	0
cell 1	0.8	0.2	0.8	0.7	0.65	3.1	7.2
cell 2	0.45	0.55	0.2	0.35	0.2	0.5	1.2

$$\mathcal{L}_{cls} = (1 - 0.8)^2 + (0 - 0.2)^2$$

3 – VOC Dataset

Total Loss Function

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{cls}}$$

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

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

❸ $$\mathcal{L}_{\text{obj}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} (C_{ij} - \hat{C}_{ij})^2$$

❹ $$+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{noobj}} (C_{ij} - \hat{C}_{ij})^2$$

❺ $$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{S^2} 1_{i}^{\text{obj}} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2$$

1. penalize bad localization of center of cells
2. penalize the bounding box with inaccurate height and width. The square root is present so that errors in small bounding boxes are more penalizing than errors in big bounding boxes.
3. make confidence score equal to the IoU between the object and the prediction when there is one object
4. make confidence score close to 0 when there are no object in the cell
5. a simple classification loss (not explained in the article)

AI VIET NAM

@aivietnam.edu.vn

Thanks!

Any questions?