In511 Commandes Intelligentes

The second part of the course, October 2025

by Aybüke ÖZTÜRK SURI

The objective of the second part of the course

How to do object detection by training your own datasets and show the object on the camera by using vocal commands.

IPSA
ÉCOLE D'INGÉNIEURS

An ongoing project: Robot Voya

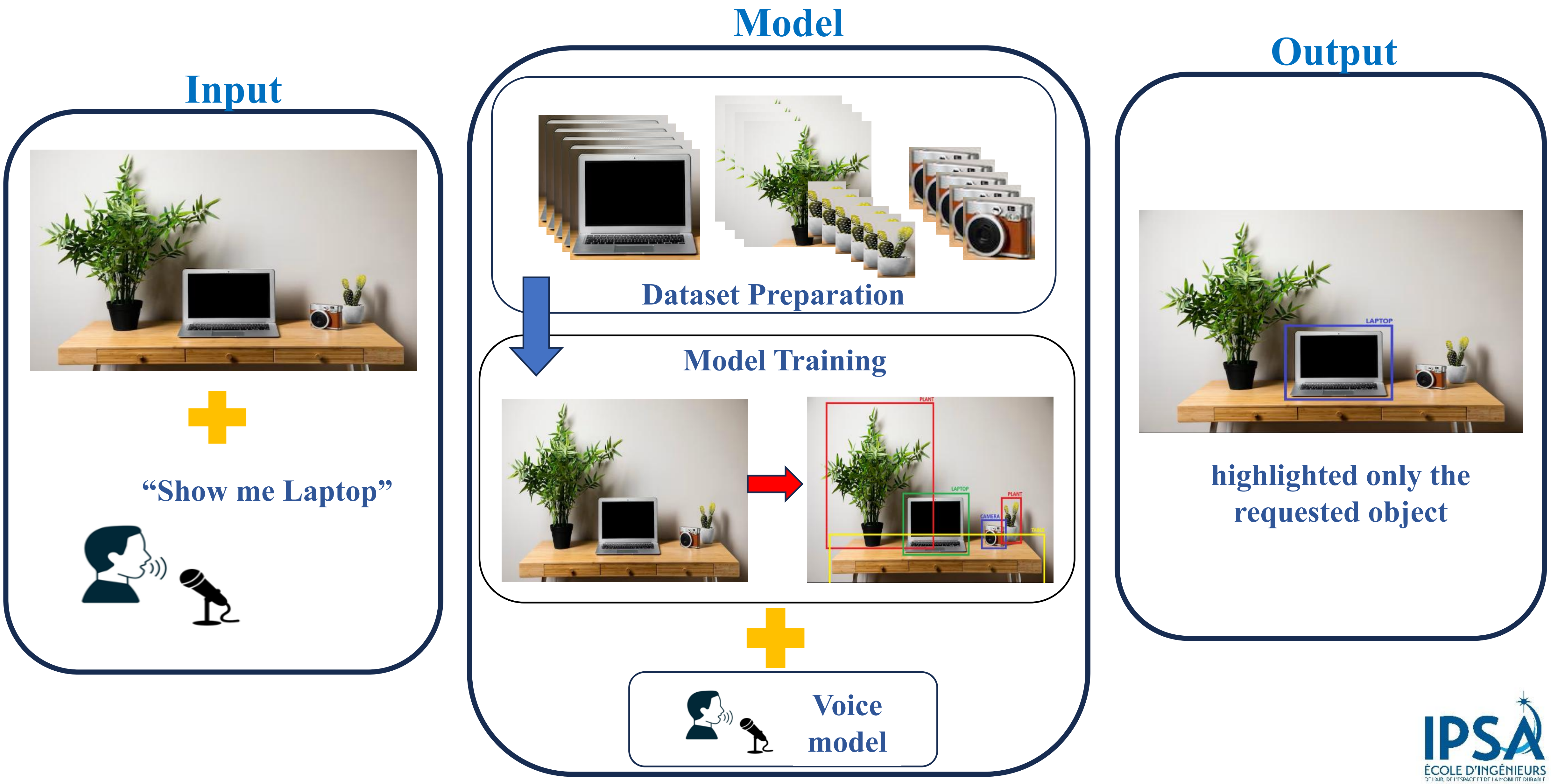
- Autonomous Movement
  - Ability to navigate environments using sensors (LIDAR, cameras).
  - Path planning and obstacle avoidance
- Perception System
  - AI-based vision for object classification
- Manipulation
  - Pick and place objects
- Human interaction
  - Human-tracking system with vision or sensors
  - Gesture recognition to interpret simple commands (e.g., waving to stop or follow)

IPSA
ÉCOLE D'INGÉNIEURS

Data Collection & Preparation

104 object to use in your experiments

Voice-Activated Real-Time Object Detection Using Custom Datasets



Data Collection & Preparation

- Create the data folder

In511/custom_dataset

Name	Type
data	Fichier source Yaml
images	File folder
labels	File folder

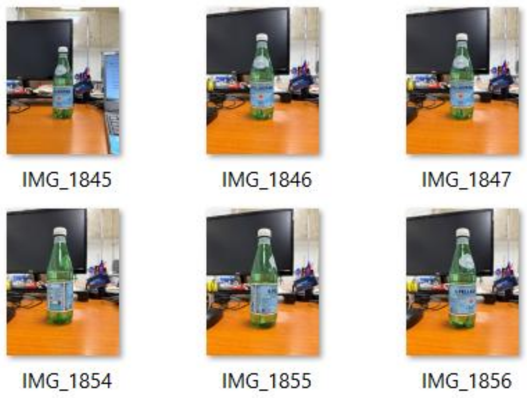
```
# data.yaml
path: ./custom_office_dataset  # dataset root dir
train: images/train  # train images (relative to path)
val: images/val  # val images (relative to path)

# Classes
nc: 3  # number of classes
names: ['Bottle', 'Bee', 'Stapler']
```

In511/custom_dataset/images

Name	Date modified
train	7/18/2025 11:31 AM
val	7/18/2025 12:02 PM

In511/custom_dataset/images/train &
In511/custom_dataset/images/val



your custom images

Data Collection & Preparation

- Create the data folder

In511/custom_dataset

Name	Type
data	Fichier source Yaml
images	File folder
labels	File folder

In511/custom_dataset/label

Name	Date modified
train	7/18/2025 11:31 AM
val	7/18/2025 12:02 PM

In511/custom_dataset/labels/train

Name	Type
classes	Document texte
IMG_1845	Document texte
IMG_1846	Document texte
IMG_1847	Document texte
IMG_1848	Document texte
IMG_1849	Document texte

classes.txt

Fichier Modifier Affichage

```
Bottle
Bee
Stapler
```

your custom object names

Data Collection & Preparation

- Create the data folder

In511/custom_dataset

Name	Type
data	Fichier source Yaml
images	File folder
labels	File folder

In511/custom_dataset/label

Name	Date modified
train	7/18/2025 11:31 AM
val	7/18/2025 12:02 PM

In511/custom_dataset/labels/train

Name	Type
classes	Document texte
IMG_1845	Document texte
IMG_1846	Document texte
IMG_1847	Document texte
IMG_1848	Document texte
IMG_1849	Document texte

IMG_1845.txt

Fichier Modifier Affichage

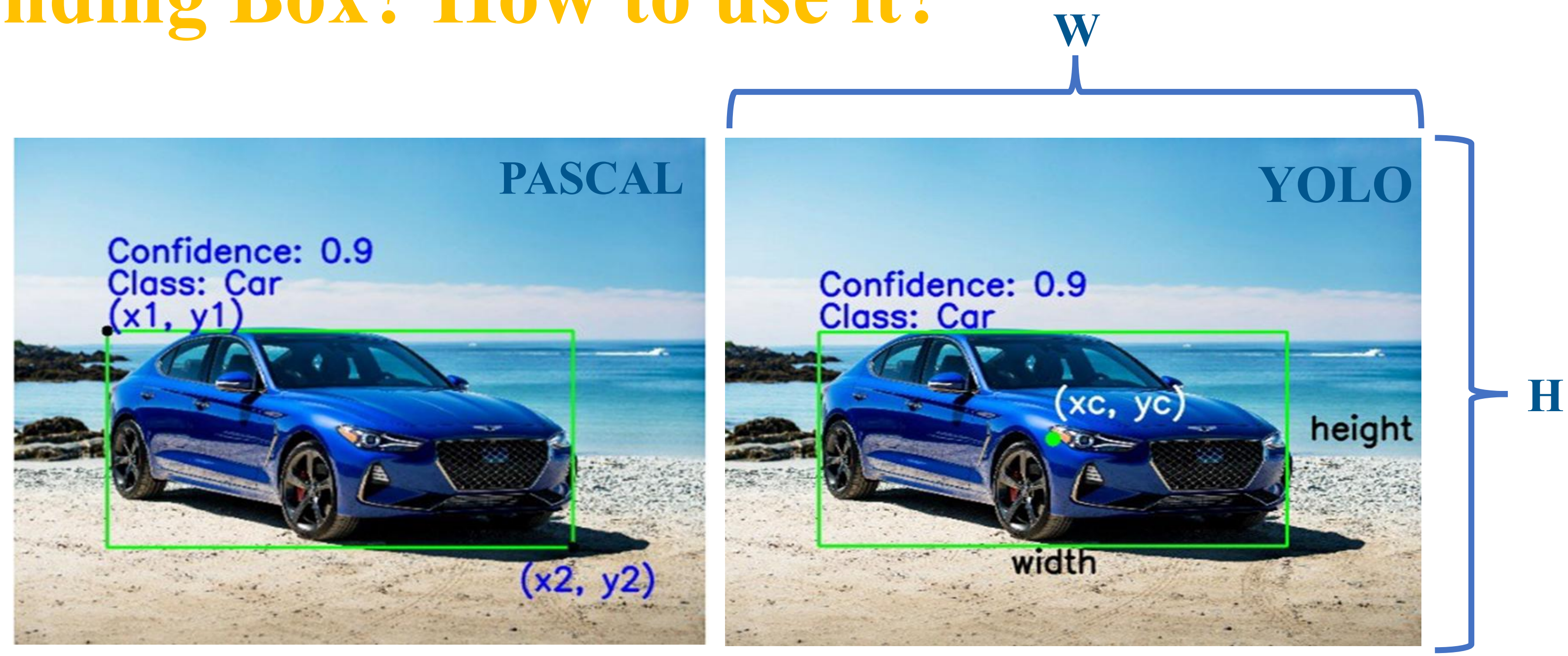
```
0 0.502801 0.512430 0.188142 0.500000
```

Class ID

Bounding Box Values

IPSA ÉCOLE D'INGÉNIEURS

What is Bounding Box? How to use it?



https://nanonets.com/blog/image-processing-and-bounding-boxes-for-ocr/

Format	Coordinates	Storage	Normalization
PASCAL VOC	xmin, ymin, xmax, ymax	XML	No (pixels)
COCO	x, y, width, height	JSON	No (pixels)
YOLO	x_center, y_center, width, height	TXT	Yes (0–1)

the center of the bounding box in pixels:

$$X_{center} = \frac{X_1 + X_2}{2} \qquad Y_{center} = \frac{Y_1 + Y_2}{2}$$

the normalized center coordinates:

$$x_c = \frac{X_{\mathrm{center}}}{W} \qquad y_c = \frac{Y_{\mathrm{center}}}{\mathrm{H}}$$

W x H: the size of the Image, **width:** the width of the bounding box, **height:** the height of the bounding box.

$(\boldsymbol{X_1}, \boldsymbol{Y_1})$**:** the X and Y coordinates of the top left corner of the rectangle.

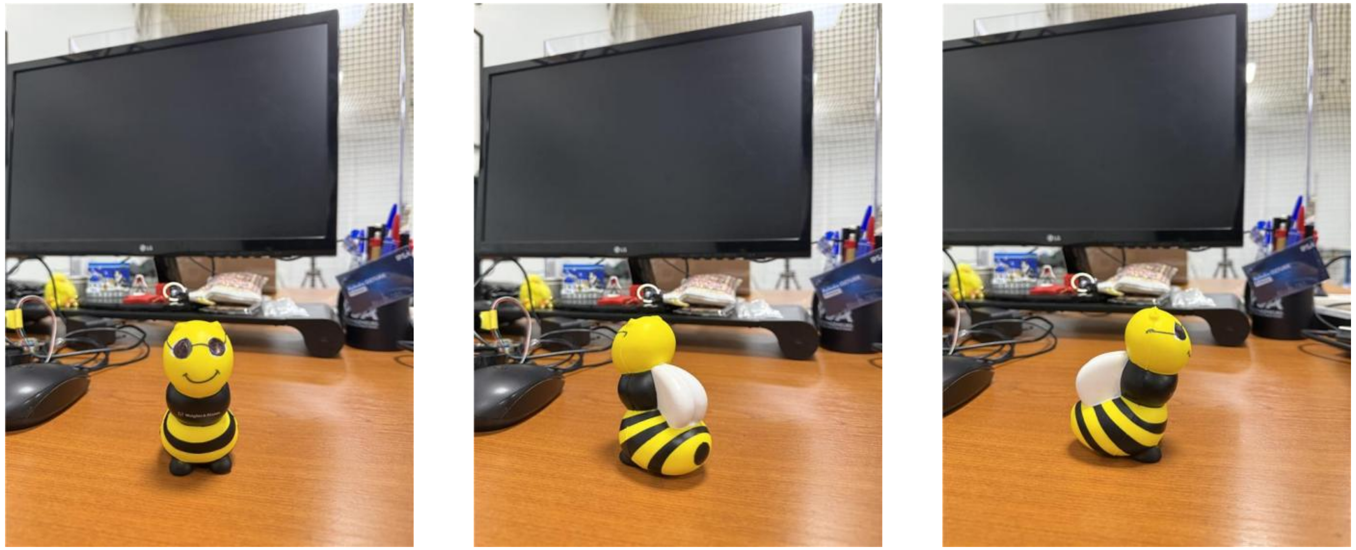
$(\boldsymbol{X_2}, \boldsymbol{Y_2})$**:** the X and Y coordinates of the bottom right corner of the rectangle.

$(\boldsymbol{X_c}, \boldsymbol{Y_c})$**:** the normalized coordinates of the center of the bounding box.

Data Collection & Preparation

Capture **100 images** of the custom object using your phone.

- **75 images** will be used for training.
- **25 images** will be used for validation.
- Make sure to capture the object from **different angles** and at **various distances** to improve dataset diversity.

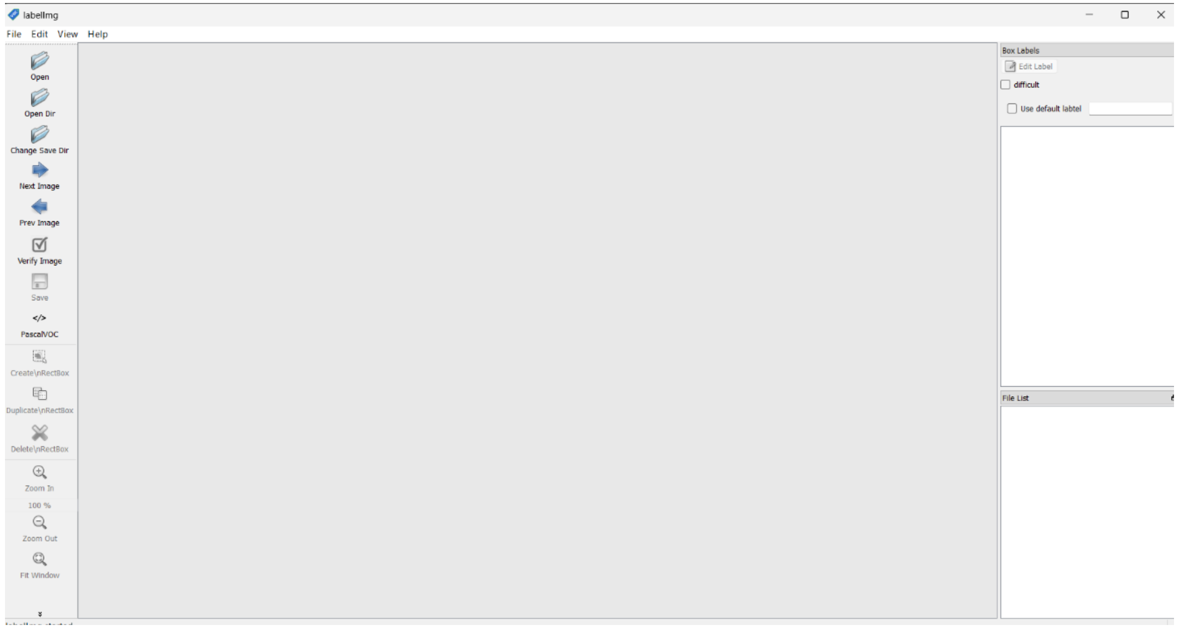


Example images of a custom object

Data Collection & Preparation

- If your images are in **HEIC format**, convert them to **PNG format** using the provided example code on Moodle: **heicToPngConverter.py** or you can find a heicToJpeg code online.
- Create labels for your dataset
  You can download the desktop version of the labeling tool **LabelImg** from the following link:
  https://github.com/HumanSignal/labelImg/releases

LabelImg Application

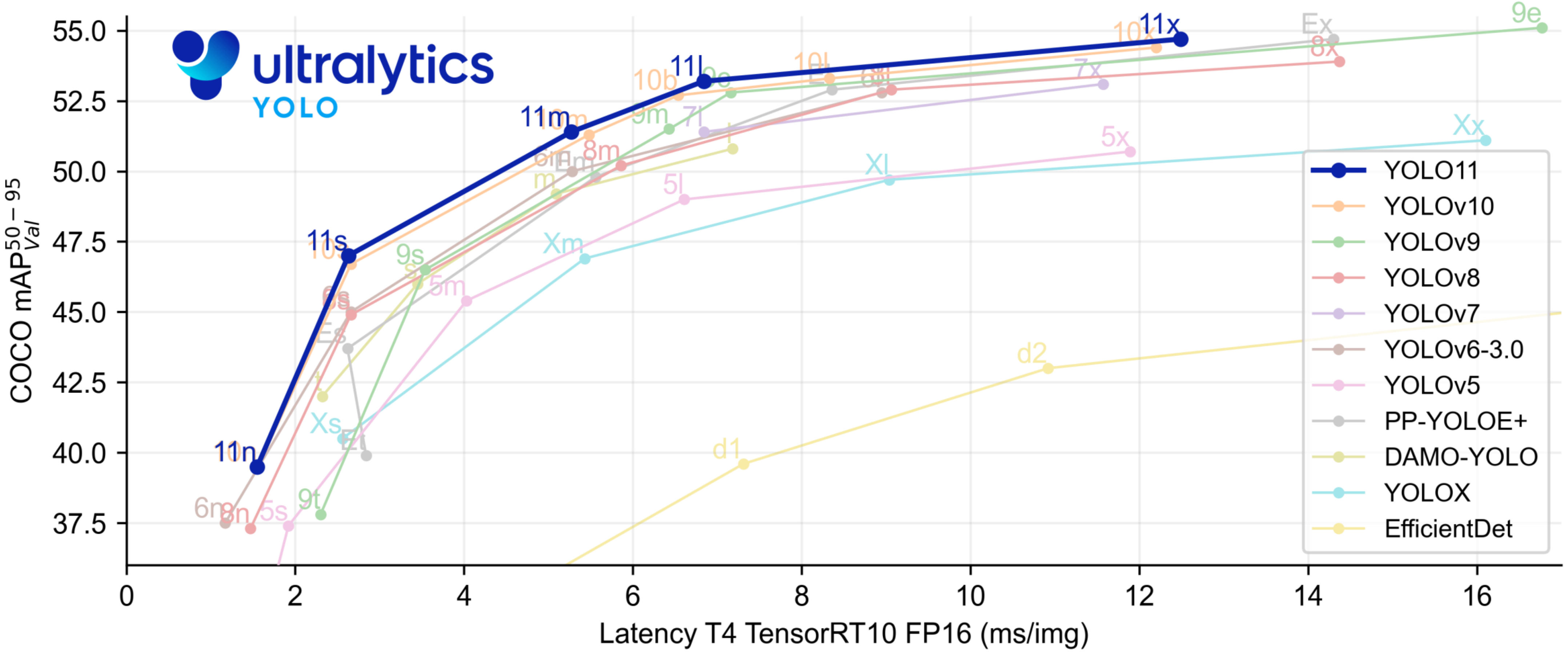


Model Training - History of YOLO

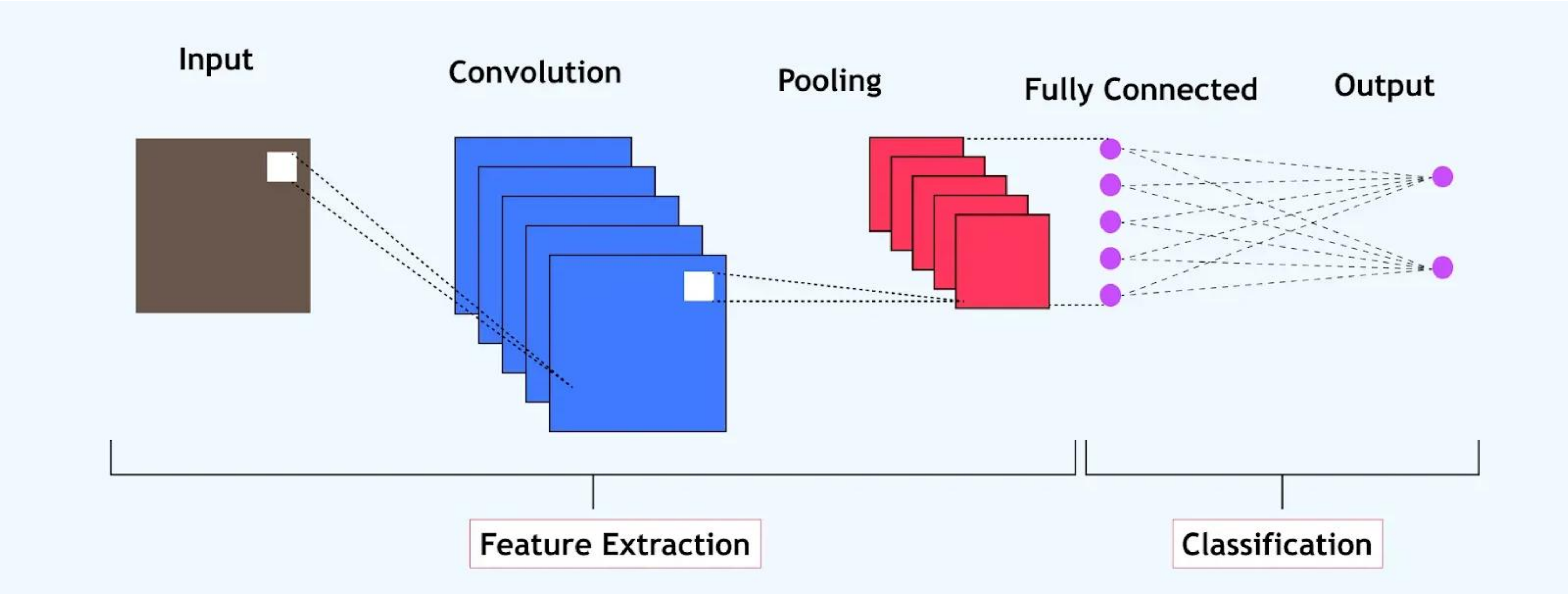
Year	Version	Highlights
2016	**YOLOv1** – *Redmon et al.*	Introduced YOLO in **“You Only Look Once: Unified, Real-Time Object Detection” (CVPR 2016)**, establishing the foundation of single-pass object detection.
2017	**YOLOv2 / YOLO9000**	Added anchor boxes, batch normalization, and real-time detection of 9,000 classes.
2018	**YOLOv3**	Adopted Darknet-53 backbone, multi-scale detection, and improvements in accuracy.
2020–2023	**YOLOv4, v5, v6, v7, v8**	Continuous improvements in speed and accuracy. e.g., YOLOv7 set new state-of-the-art (SOTA) results.
2024	**YOLOv10**	Introduced NMS-free training, dual-assignment strategy, and holistic efficiency improvements.
September 2024	**YOLO11**	Released by Ultralytics at YOLO Vision 2024 (YV24); features architectural refinements (C3k2, C2PSA blocks), fewer parameters, improved speed, and multi-task support.
February 18, 2025	**YOLOv12** – *Tian et al.*	Attention-centric architecture (Area Attention, R-ELAN, FlashAttention). Achieves SOTA mAP with very low latency—e.g., YOLOv12-N hits 40.6% mAP in **1.64 ms**, outperforming prior versions.

Reference for ULTRALYTICS: https://docs.ultralytics.com/models/yolo11/

Performance of YOLO Versions



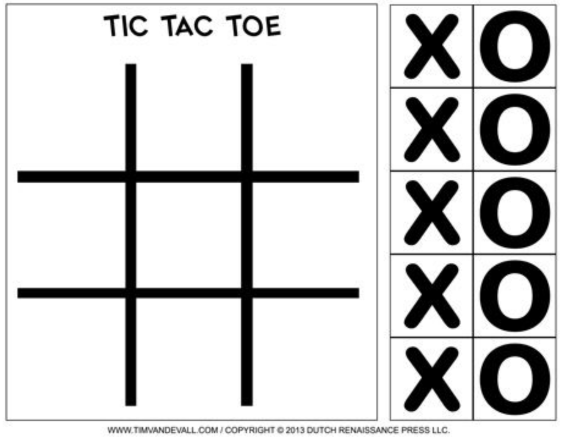
Basic Architecture of Convolutional Neural Networks (CNNs)



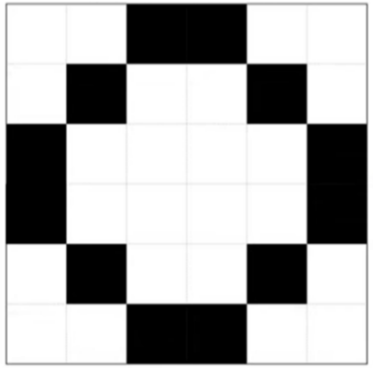
https://www.upgrad.com/blog/basic-cnn-architecture/

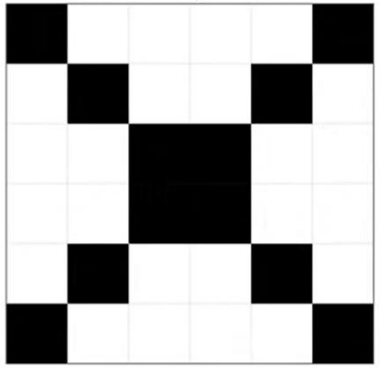
How does Convolutional Neural Networks (CNNs) work?

Example



Pixel Representation

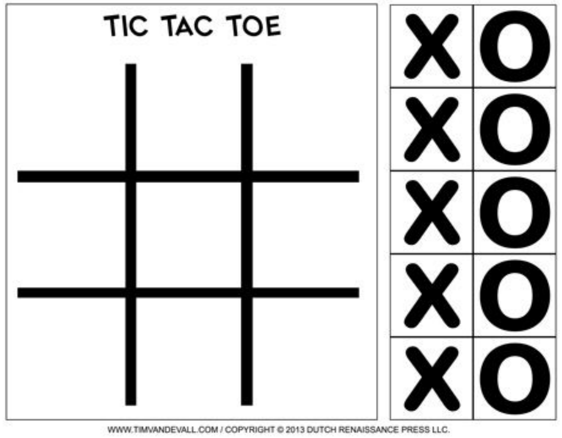




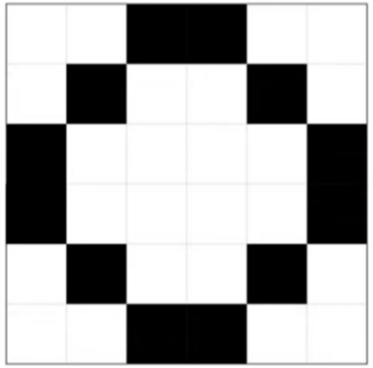
https://www.youtube.com/watch?v=HGwBXDKFk9I

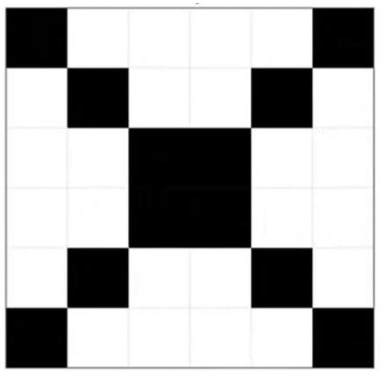
How does Convolutional Neural Networks (CNNs) work?

Example



Pixel Representation





0 for white, 1 for black pixels

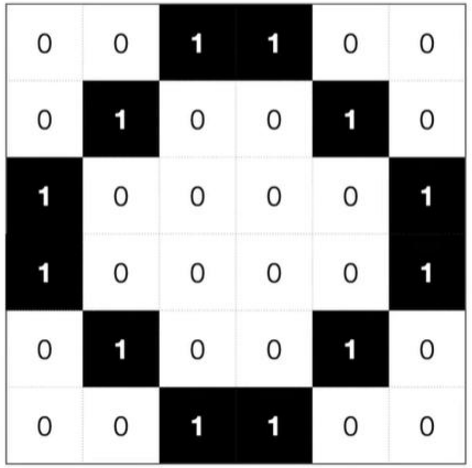
0	0	1	1	0	0
0	1	0	0	1	0
1	0	0	0	0	1
1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
0	0	1	1	0	0
0	1	0	0	1	0
1	0	0	0	0	1

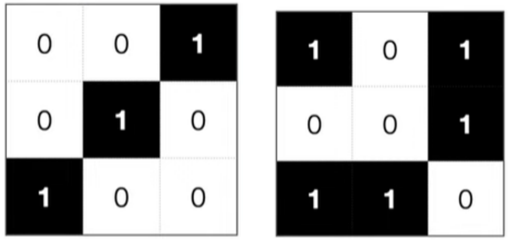
https://www.youtube.com/watch?v=HGwBXDKFk9I

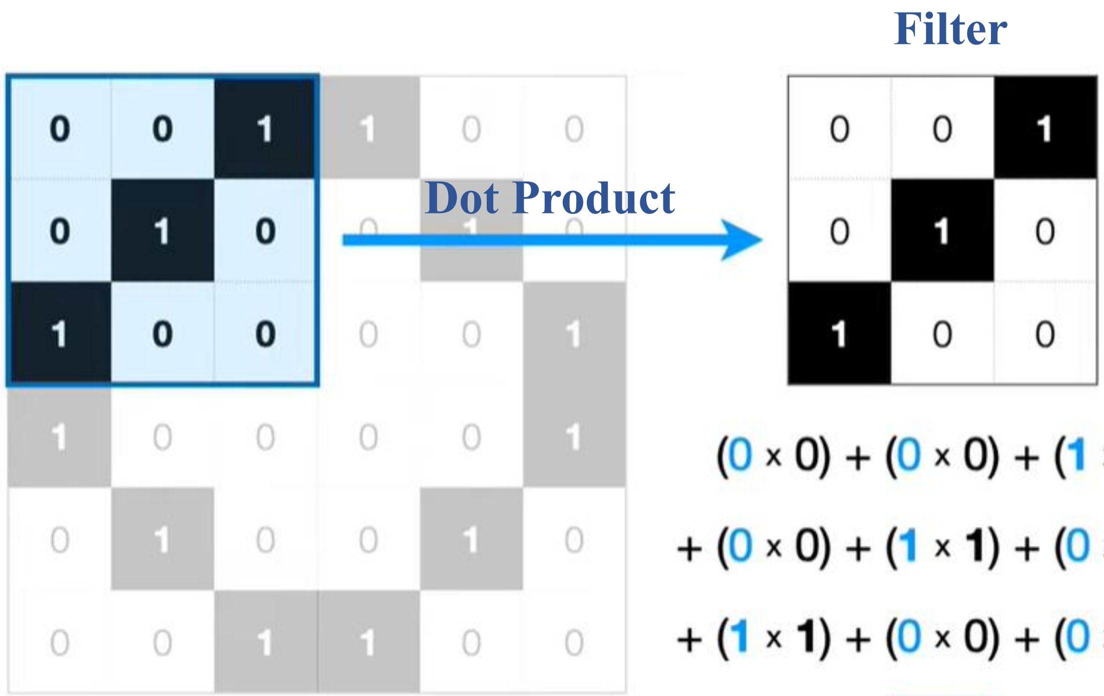
How does Convolutional Neural Networks (CNNs) work?

Example



(3x3) Kernel Matrices (Filters)





$$(0 \times 0) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (0 \times 0)$$

$$\boxed{= 3}$$

Convolutional Layer

In Python, for example;

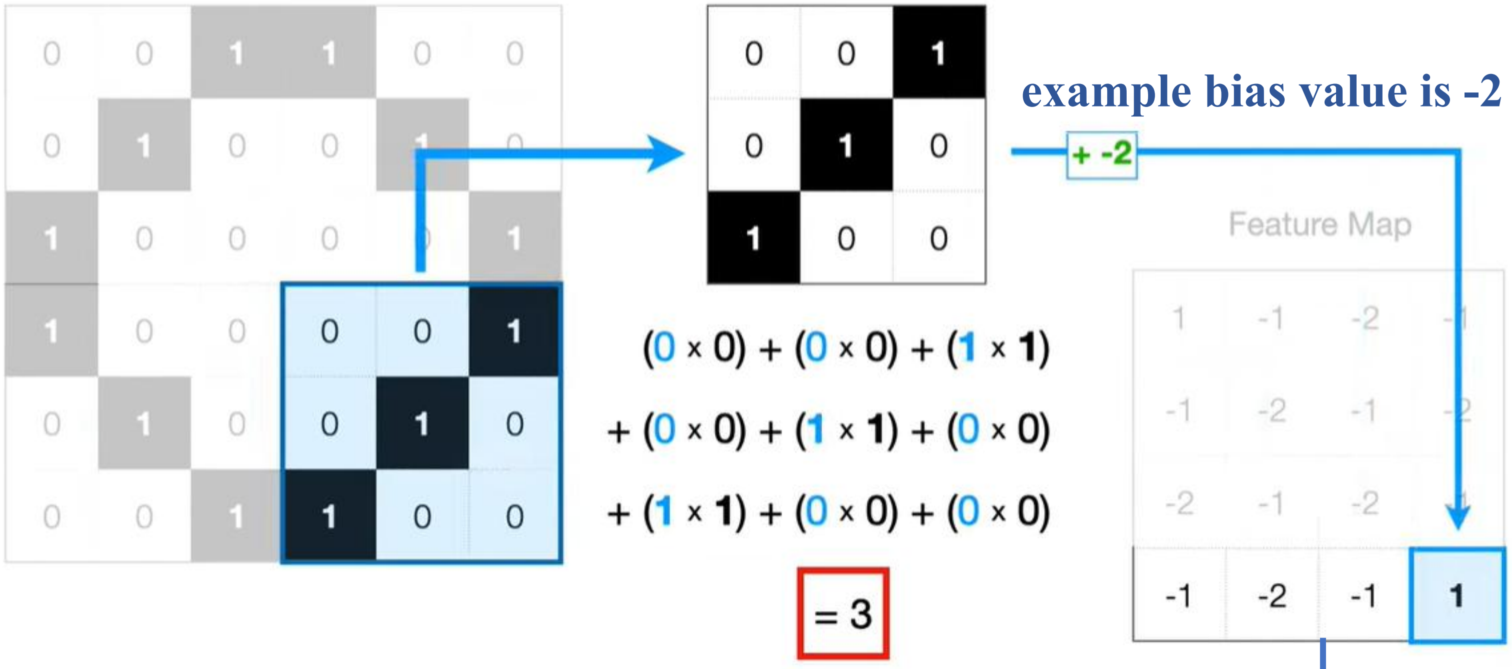
nn.Conv2d(in_channels=1, out_channels=32, kernel_size=3)

https://www.youtube.com/watch?v=HGwBXDKFk9I

How does Convolutional Neural Networks (CNNs) work?

Example

0	0	1	1	0	0
0	1	0	0	1	0
1	0	0	0	0	1
1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0



$$(0 \times 0) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (1 \times 1) + (0 \times 0) + (0 \times 0)$$

$$\boxed{= 3}$$

(3x3) Kernel Matrices (Filters)

0	0	1
0	1	0
1	0	0

1	0	1
0	0	1
1	1	0

Convolutional Layer

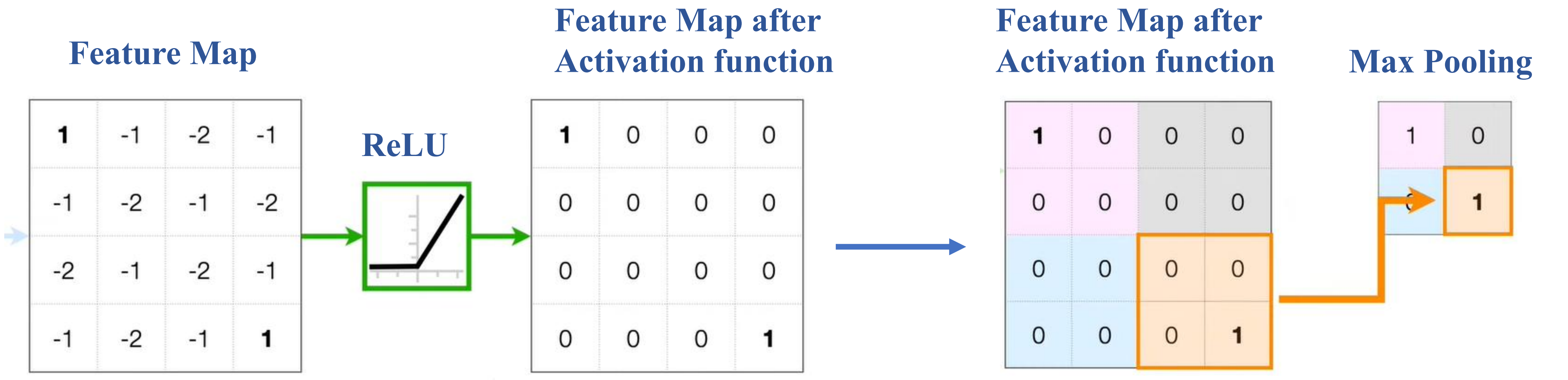
Feature Map

1	-1	-2	-1
-1	-2	-1	-2
-2	-1	-2	-1
-1	-2	-1	1

https://www.youtube.com/watch?v=HGwBXDKFk9I

How does Convolutional Neural Networks (CNNs) work?

Example



Eliminate all the values that are less than 0.

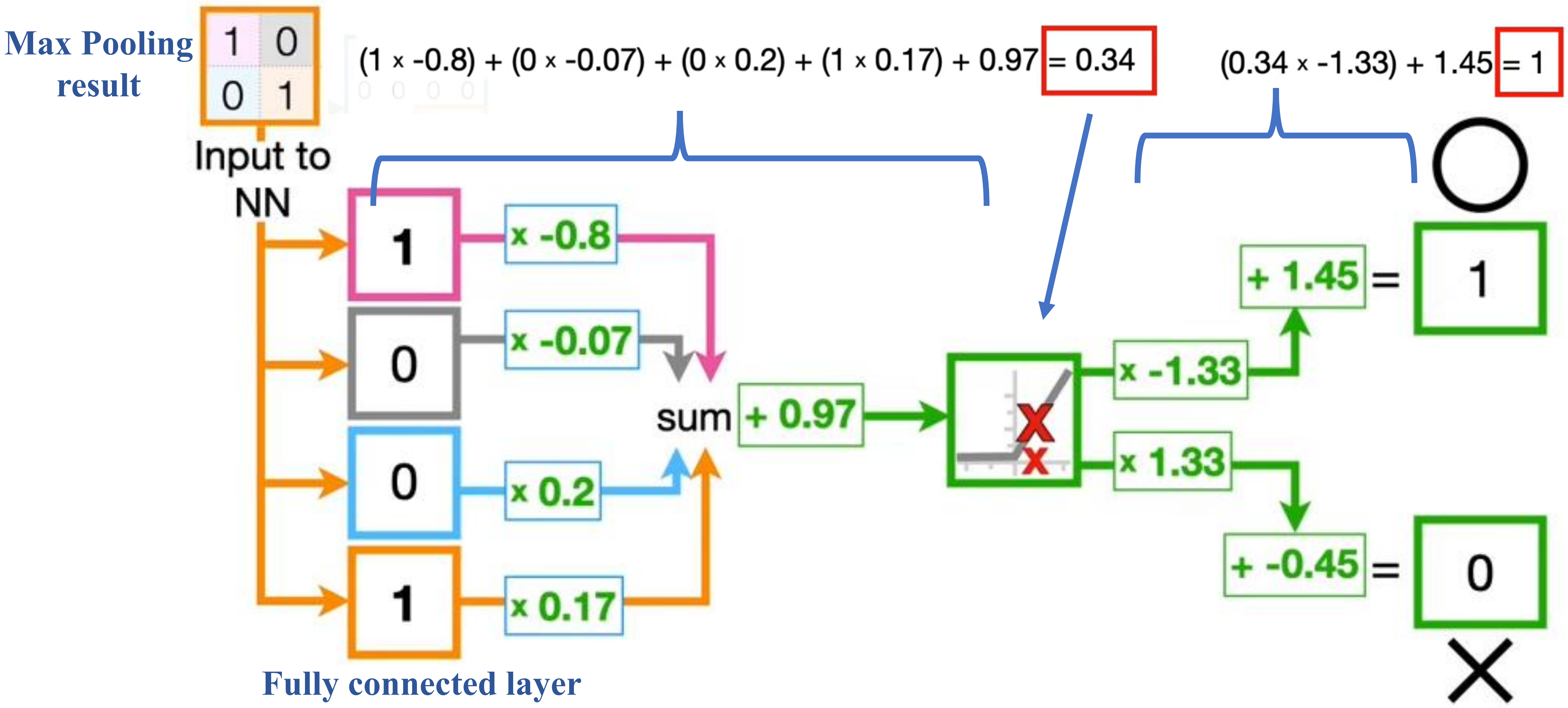
Pooling Layer

ReLU - Rectified linear unit

https://www.youtube.com/watch?v=HGwBXDKFk9I

How does Convolutional Neural Networks (CNNs) work?

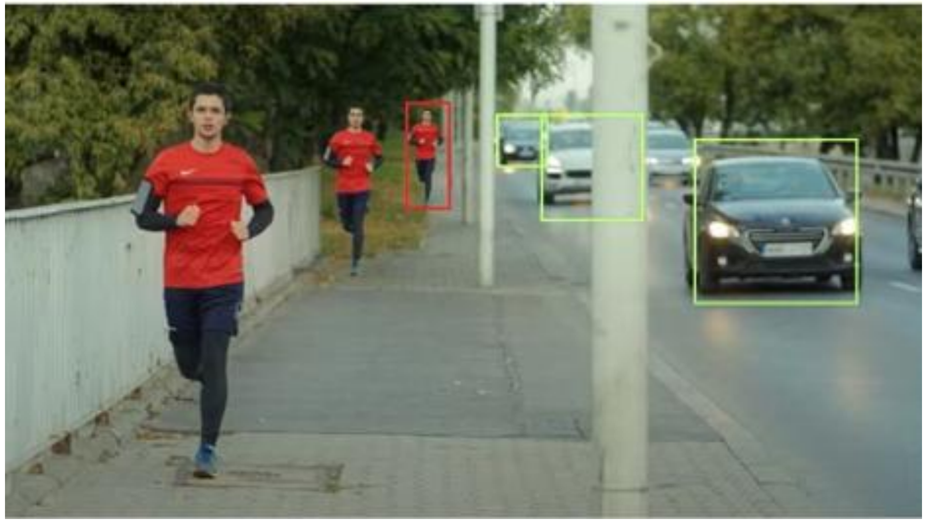
Example



https://www.youtube.com/watch?v=HGwBXDKFk9I

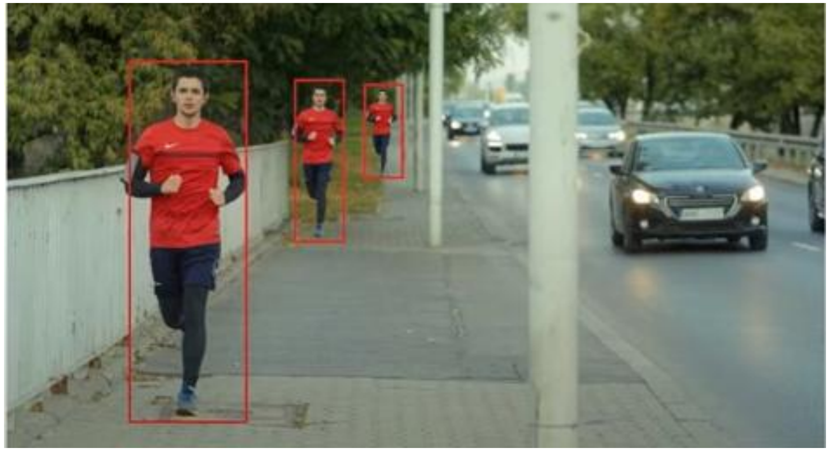
How does Region-based CNN (R-CNN) work?

Example Image

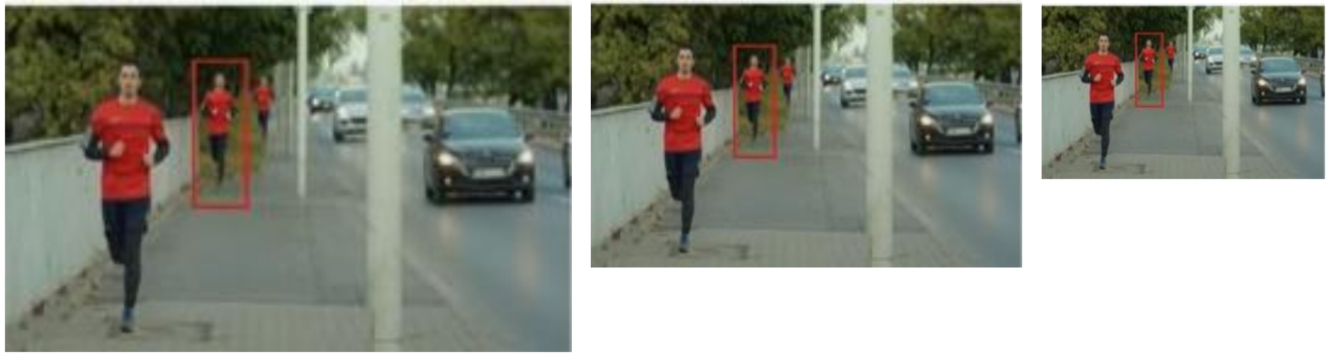


How to handle the same object with different sizes?

Crops of different size



Crops of downscale version of the image



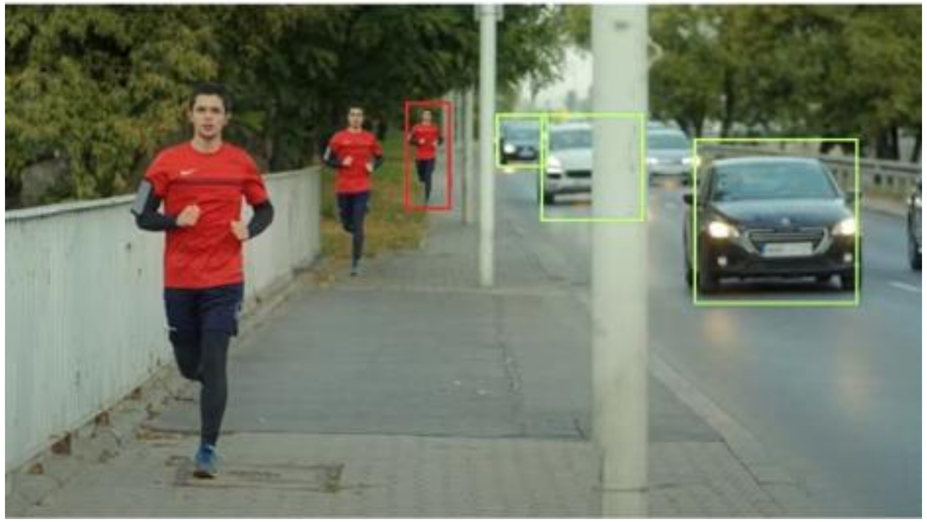
https://arxiv.org/pdf/1311.2524

https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s

https://www.youtube.com/watch?v=5DvljLV4S1E

How does Region-based CNN (R-CNN) work?

Example Image

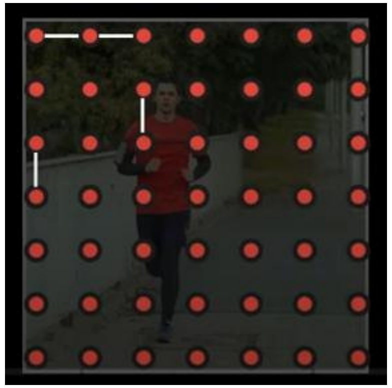


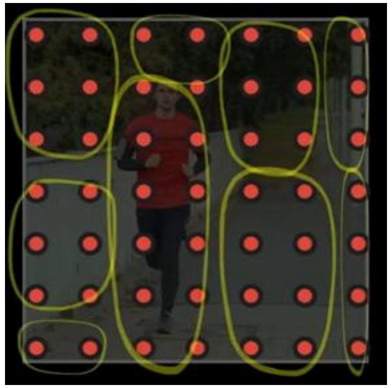
How to choose regions to pass through the CNN?

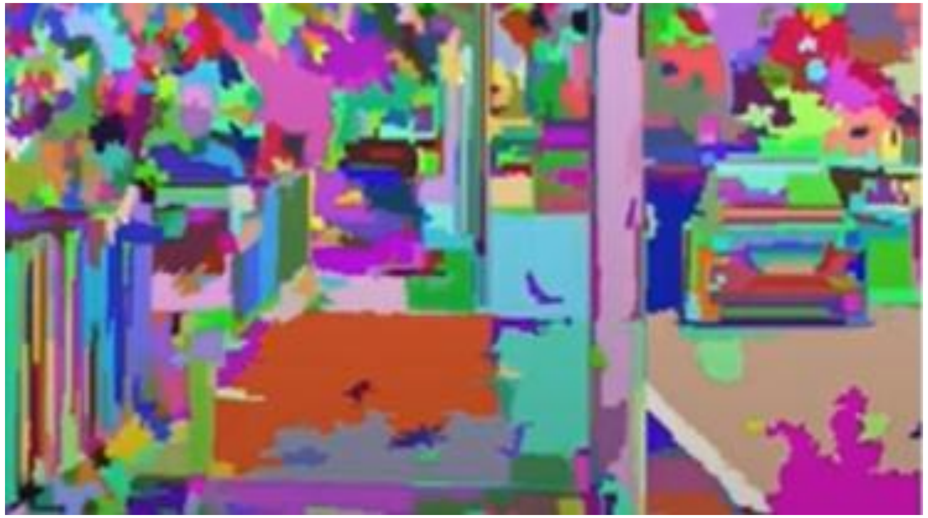
Selective Search

Step 1: segment image into regions
Step 2: merge similar regions to create larger regions

Step 1
check distance between pixels



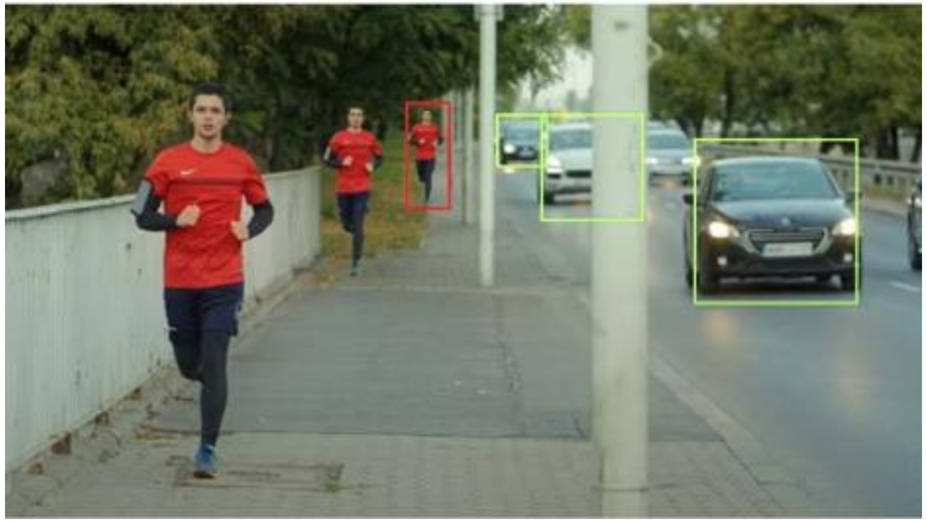




https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s https://www.youtube.com/watch?v=5DvljLV4S1E

How does Region-based CNN (R-CNN) work?

Example Image

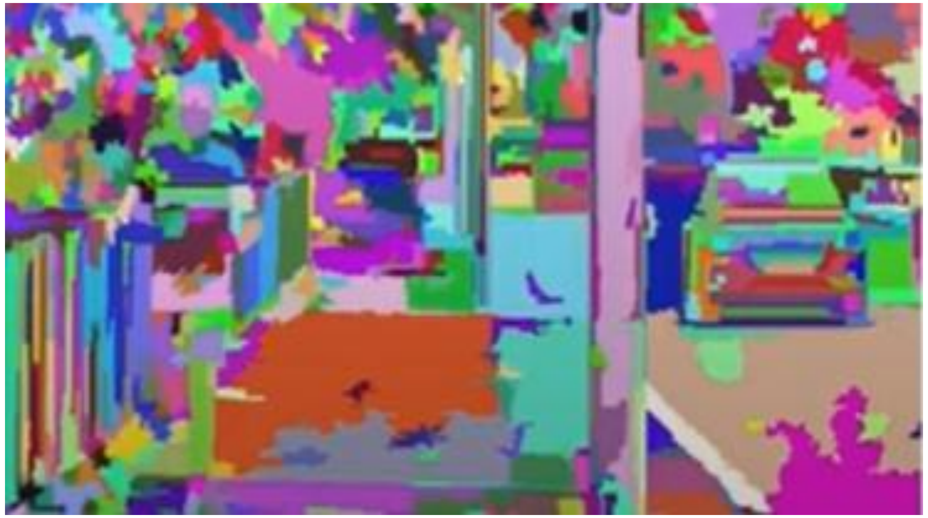


How to choose regions to pass through the CNN?

Selective Search

Step 1: segment image into regions
Step 2: merge similar regions to create larger regions

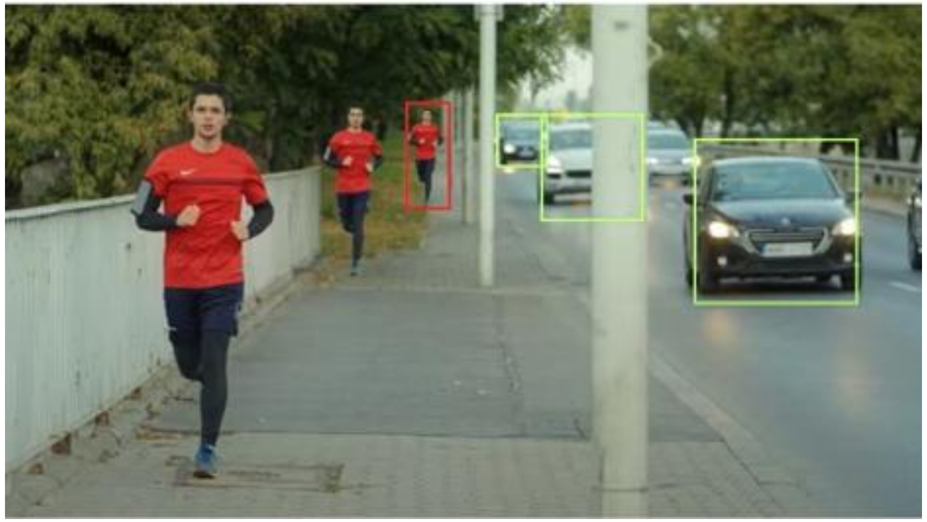
Step 2
check similarity based on color, texture, size, shape



https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s https://www.youtube.com/watch?v=5DvljLV4S1E

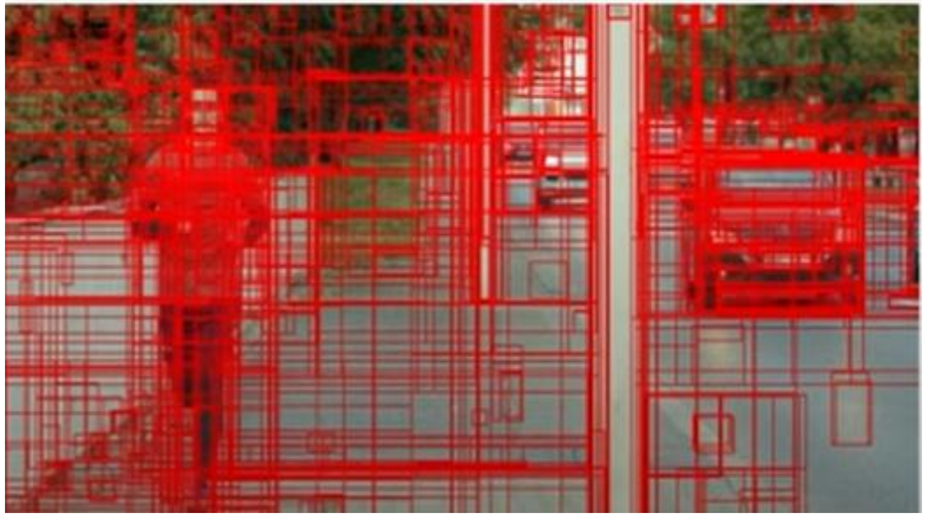
How does Region-based CNN (R-CNN) work?

Example Image

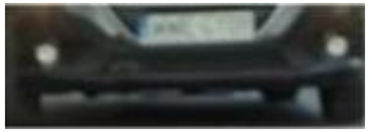


Any arbitrary shape proposed regions, how to pass them through the CNN?

Fix size region should be given as an input
Label for every chosen regions



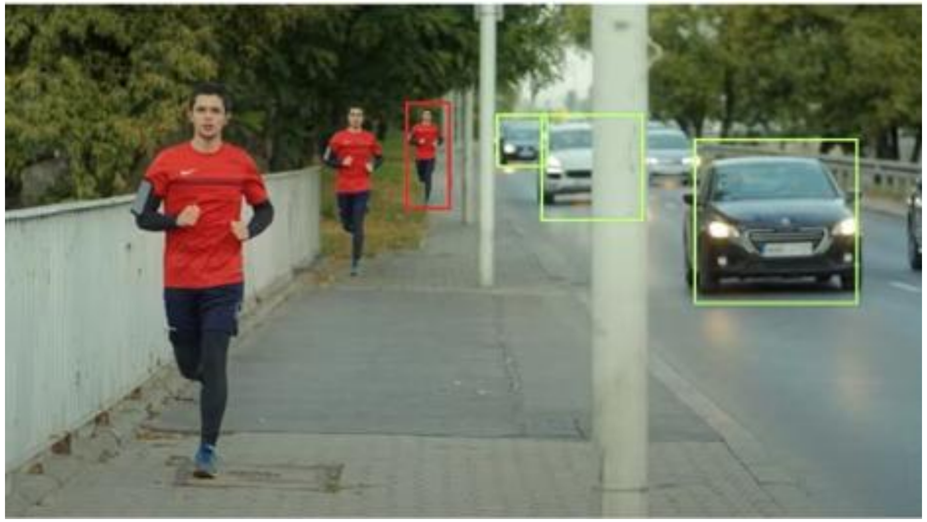
1600 region proposals

https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s https://www.youtube.com/watch?v=5DvljLV4S1E

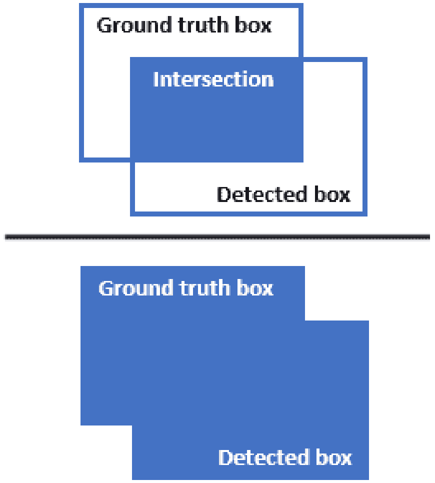
How does Region-based CNN (R-CNN) work?

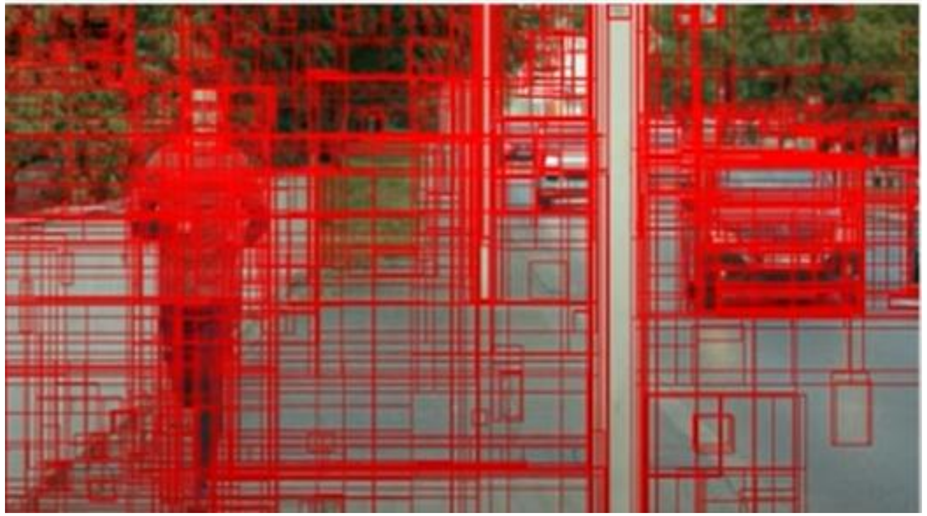
Example Image



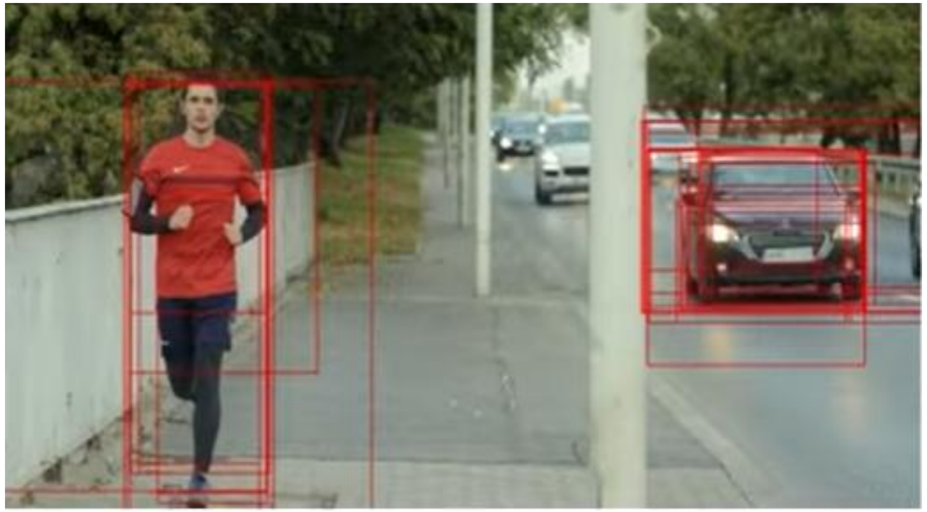
Intersection over Union

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





1600 region proposals



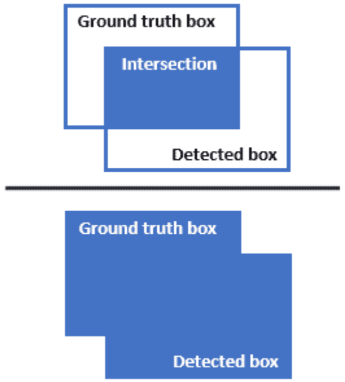
Solution: Non-maximum suppression

https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s https://www.youtube.com/watch?v=5DvljLV4S1E

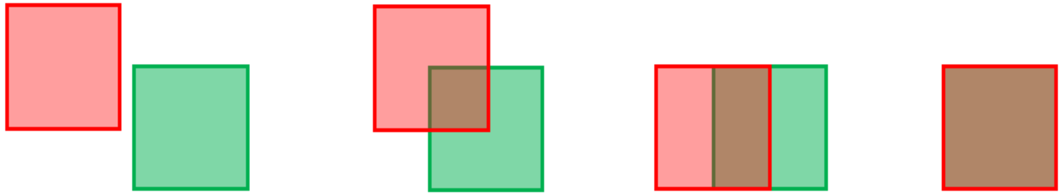
Intersection over Union (IoU)

IoU measures the accuracy of the detections

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



Examples of different IoU values



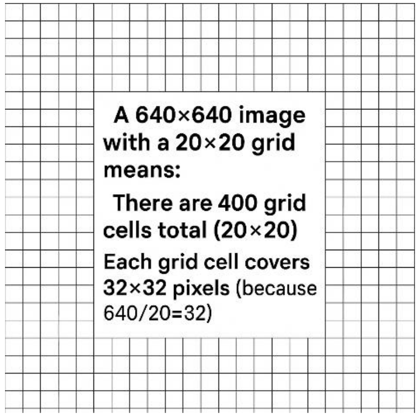
IoU = 0 IoU = 0.142 IoU = 0.333 IoU = 1

https://www.baeldung.com/cs/object-detection-intersection-vs-union

The IoU-based loss functions

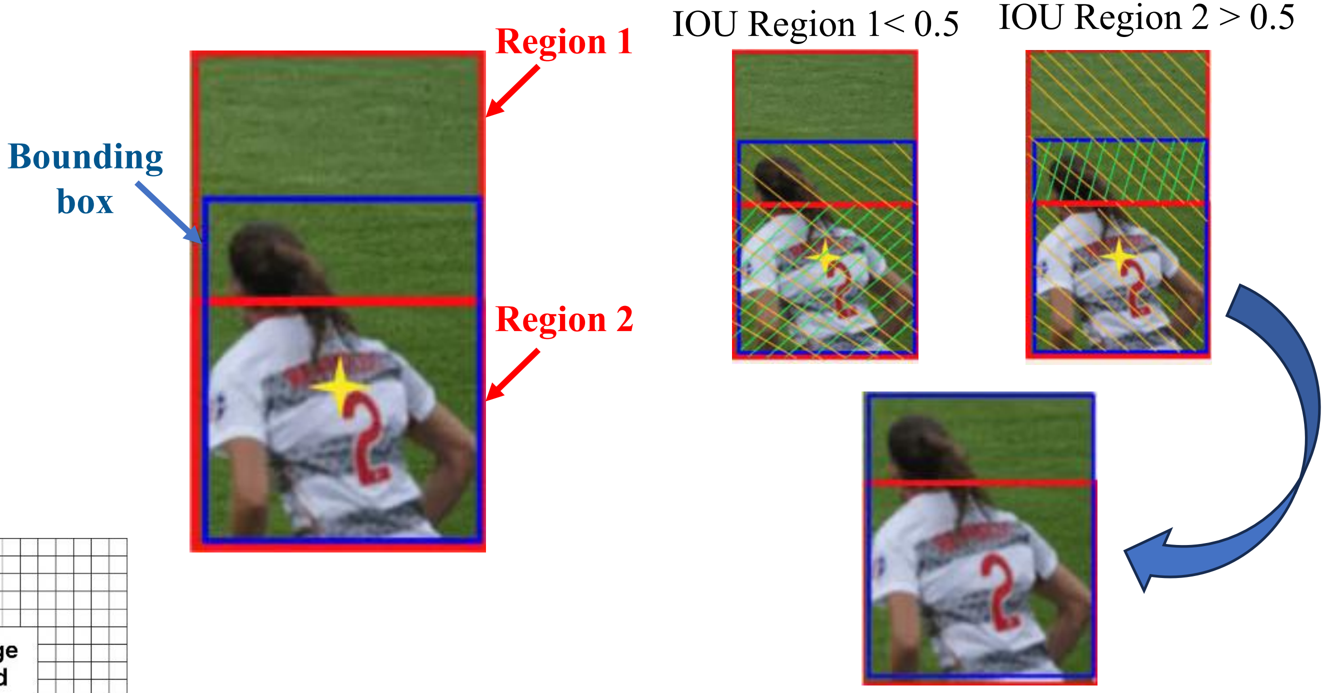
GIoU (Generalized IoU)
DIoU (Distance IoU)
CIoU (Complete IoU)
SIoU (Scalable IoU)

} **YOLOv8**



Example

- The user can define an IOU selection threshold, for instance, 0.5.
- YOLO computes the IOU of each grid cell with the formula on the left side.
- It considers only IOU > threshold.

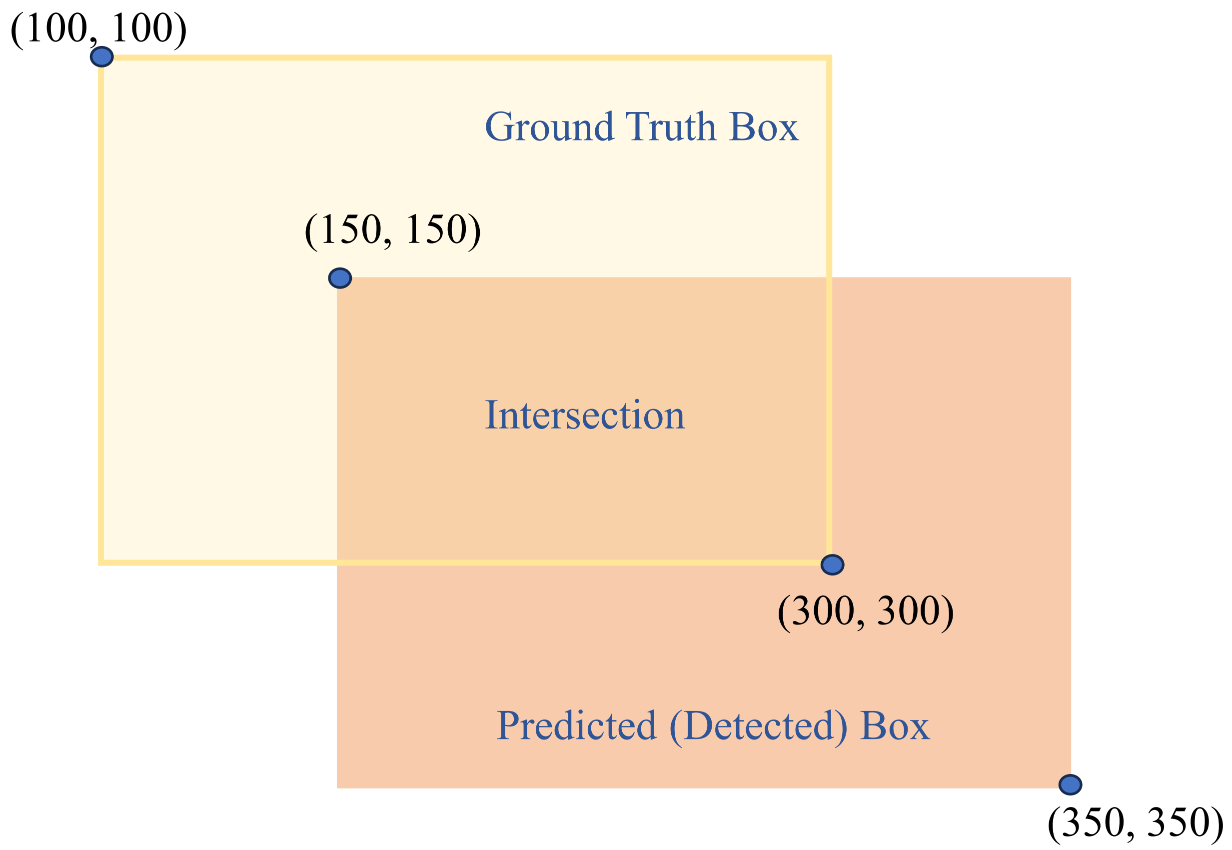


https://www.datacamp.com/blog/yolo-object-detection-explained

IPSA ÉCOLE D'INGÉNIEURS

Intersection over Union (IoU)

IoU measures the accuracy of the detections $\quad \text{IoU} = \dfrac{\textit{Area of Overlap}}{\textit{Area of Union}}$



Area of Ground Truth Box = $200 \times 200 = 40{,}000$ pixels^2

Area of Predicted Box = $200 \times 200 = 40{,}000$ pixels^2

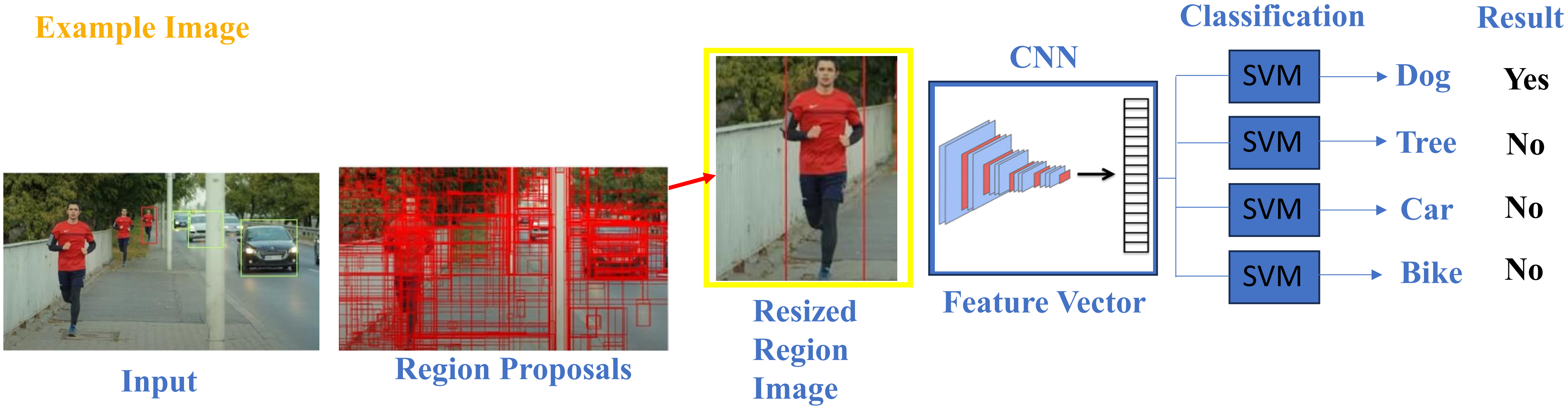
Area of Intersection = $150 \times 150 = 22{,}500$pixels^2

Union = Area(GT) + Area(P)−Intersection

$$\text{Union} = 40{,}000 + 40{,}000 - 22{,}500 = 57{,}500$$

IoU= Intersection over Union $= \dfrac{22{,}500}{57{,}500} \approx 0.391$

How does Region-based CNN (R-CNN) work?



Train with IoU >= 0.5
with ground truth
boxes of that class

SVM – Support vector machine – classification method

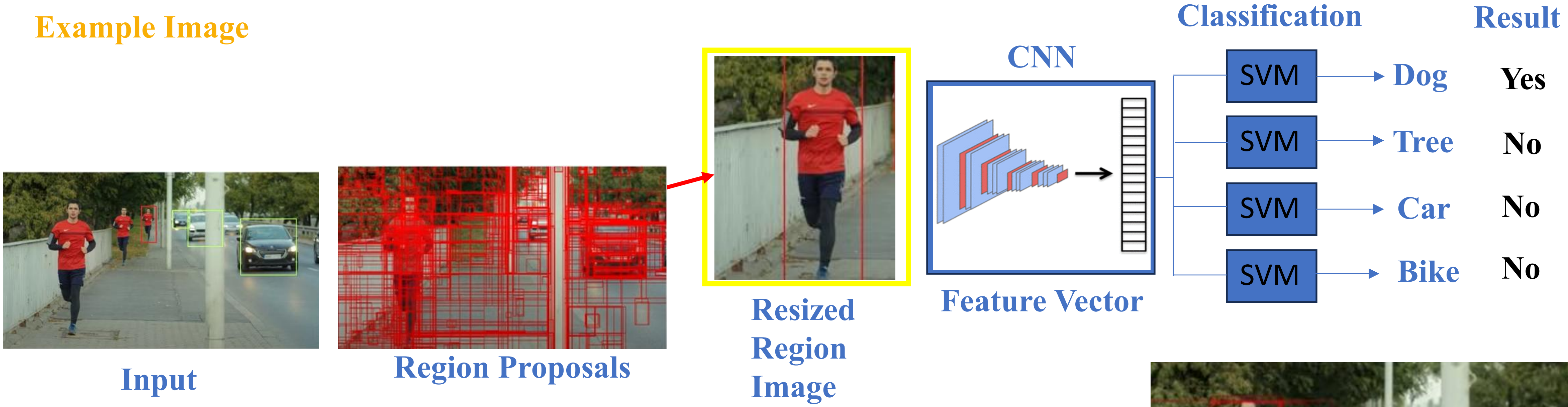
Girshick et. al (2014) “Rich feature hierarchies for accurate object detection and semantic segmentation”

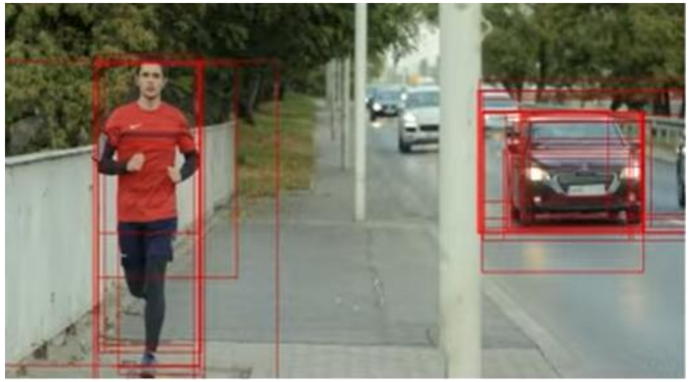
https://arxiv.org/pdf/1311.2524

Source: Redmon and https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does Region-based CNN (R-CNN) work?

Example Image





How to evaluate the model R-CNN?

- apply **Non-Maximum Suppression (NMS)** to remove duplicate detections.
- evaluate predictions using **Average Precision (AP)** per class and **Mean Average Precision (mAP)** across classes.

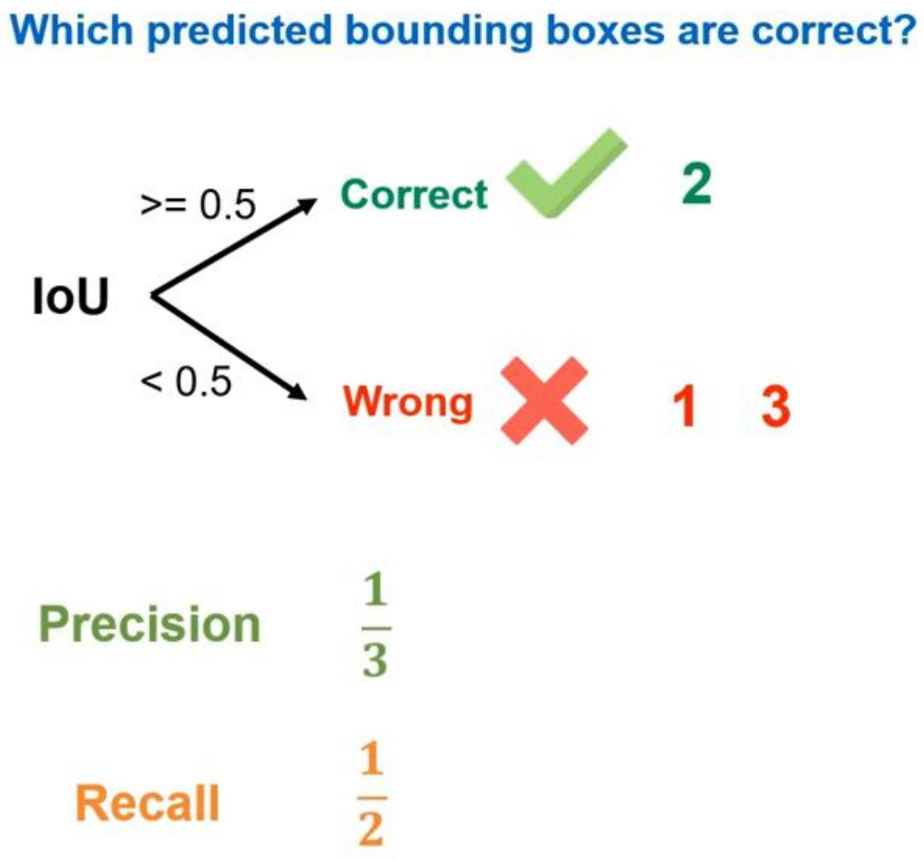
$$If\ IoU\ (P_1, P_2) > Threshold:$$

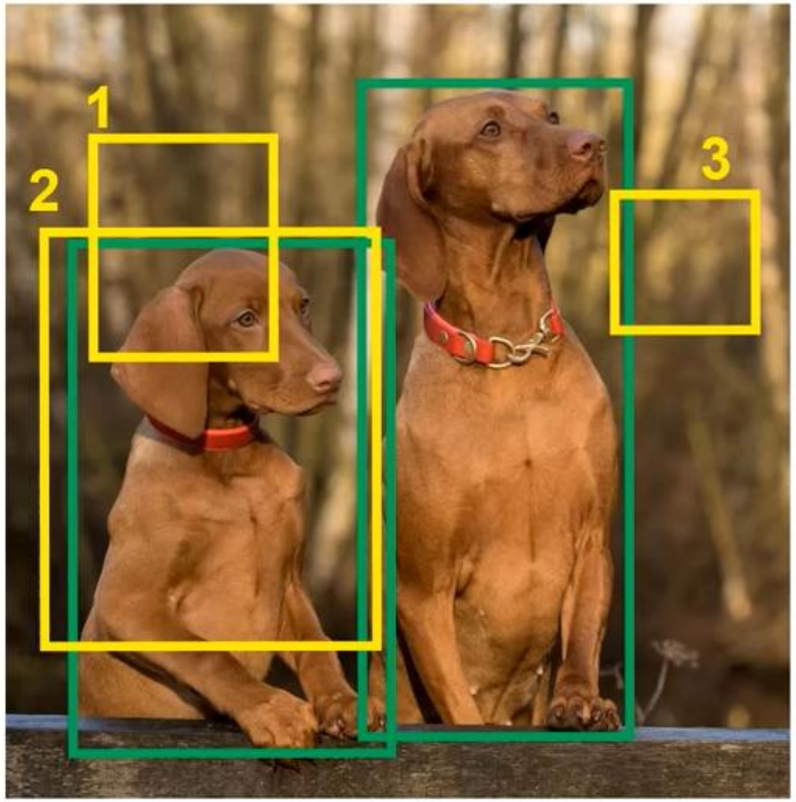
$$P = \mathrm{argmax}(C(P_1), C(P_2))$$

How does Region-based CNN (R-CNN) work?

Example Image

Mean Average Precision (mAP)



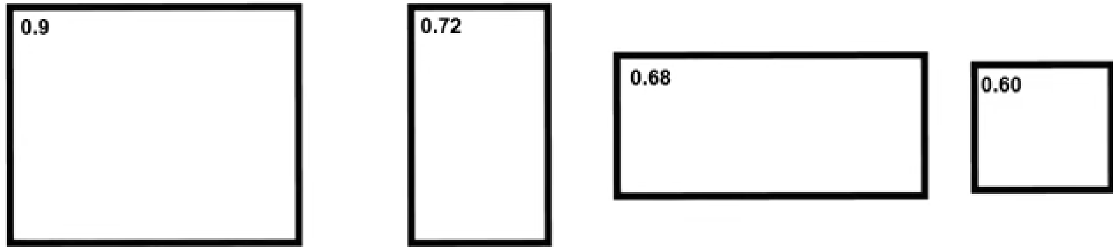


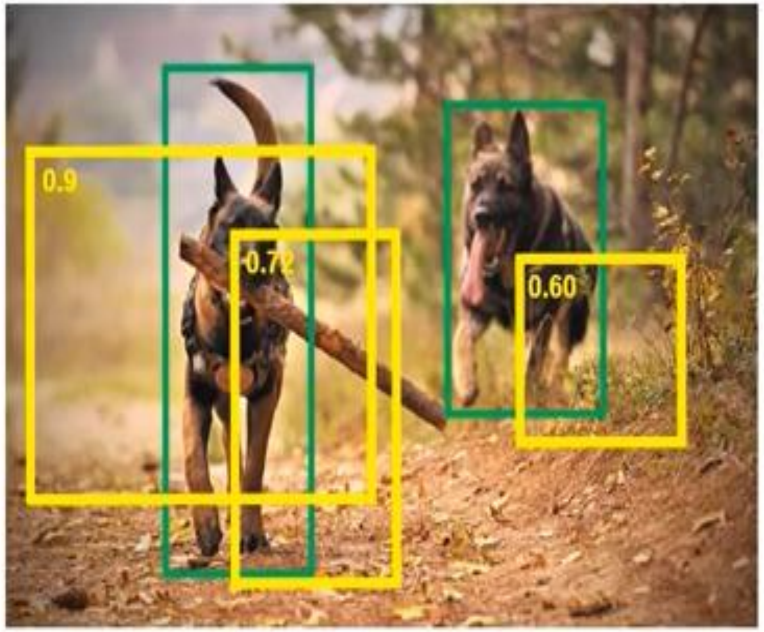
https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s

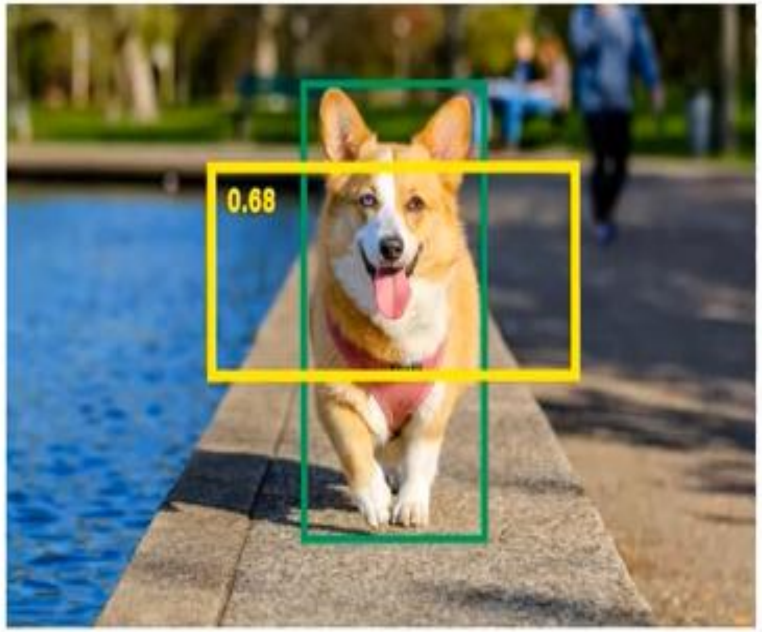
How does Region-based CNN (R-CNN) work?

Example Image

Mean Average Precision (mAP)





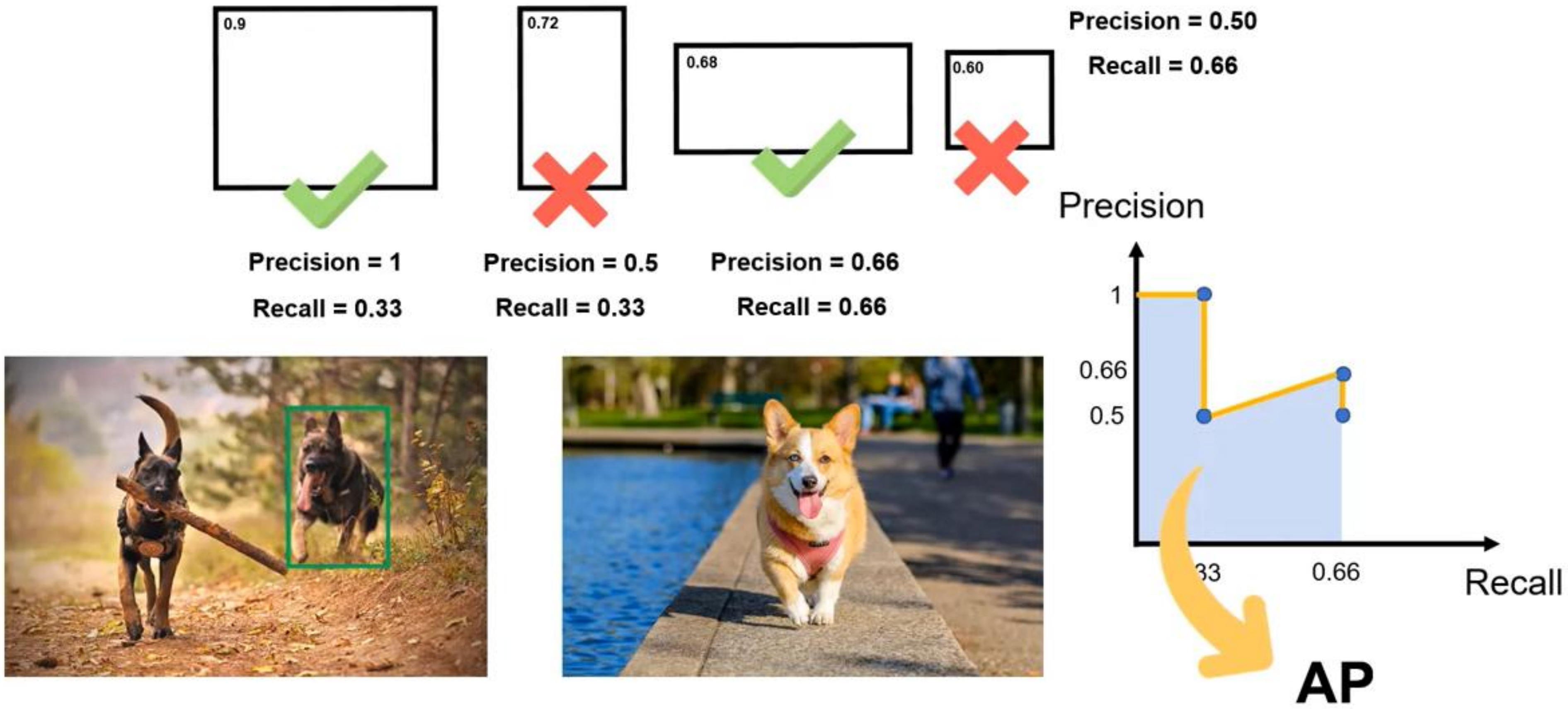


https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s

How does Region-based CNN (R-CNN) work?

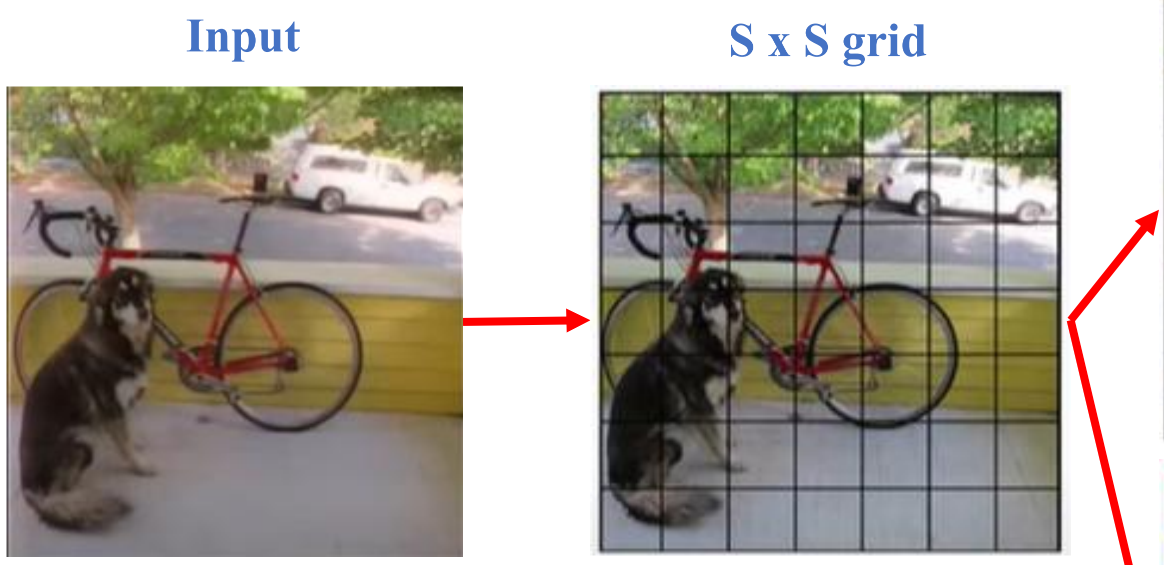
Example Image

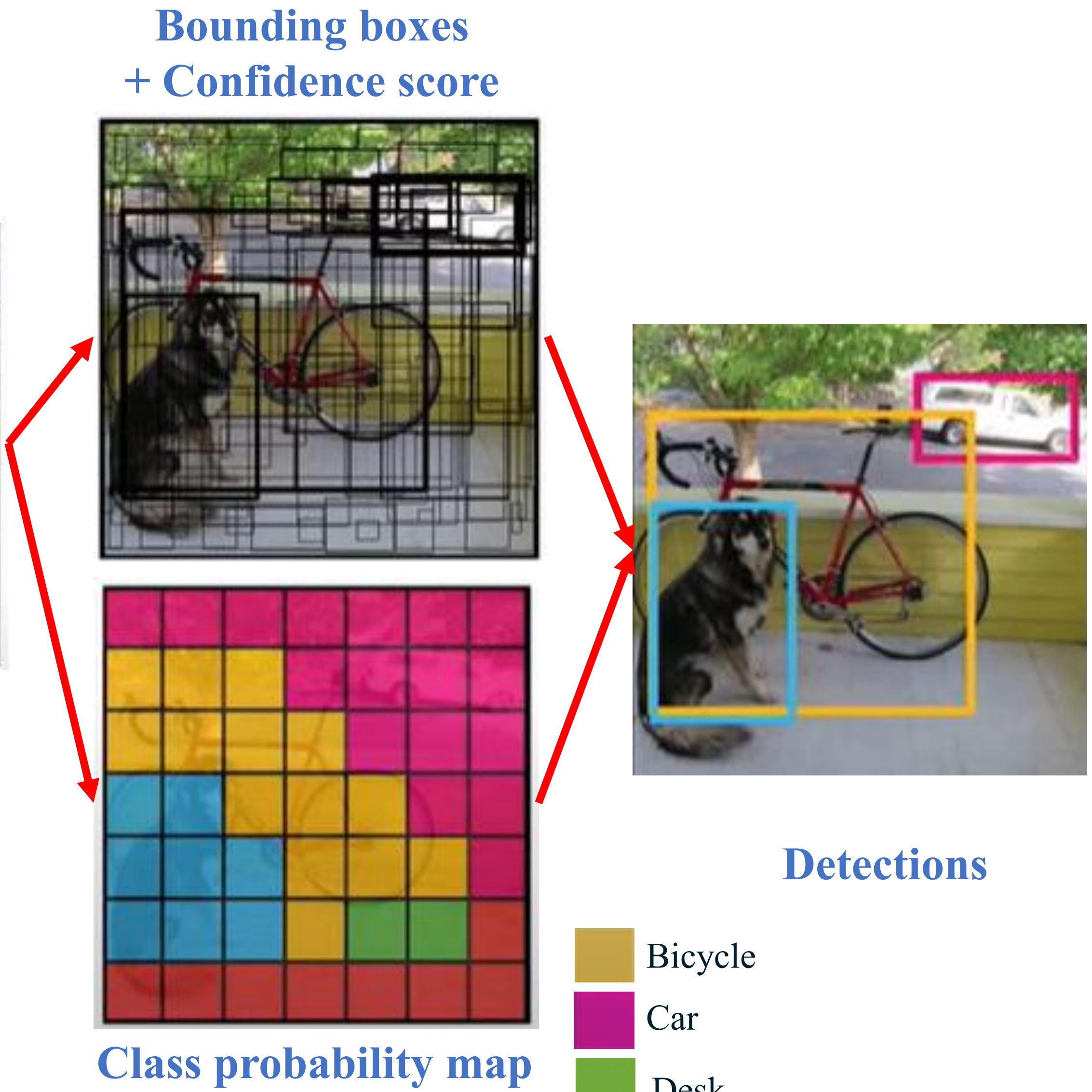
Mean Average Precision (mAP)

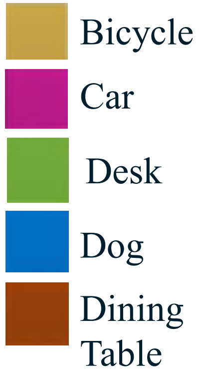


https://www.youtube.com/watch?v=nJzQDpppFj0&t=1s

How does YOLO Work?







In YOLO, each grid cell directly predicts:

- A fixed number of **bounding boxes**
- Each box's **(x, y, w, h)** coordinates
- A **confidence score & class probabilities**

Redmon et. al (2016) "you Only Look Once: Unified, Real-Time Detection"

https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf

Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does YOLO Work?

S x S grid

Output Vector Length

$B\ x\ 7 + n$

Class probabilities

$[p(c_1), p(c_2), \ldots, p(c_n)]$

Bounding box predictions

$[x_1, y_1, \sqrt{w_1}, \sqrt{h_1}, C_1]$

$[x_2, y_2, \sqrt{w_2}, \sqrt{h_2}, C_2]$

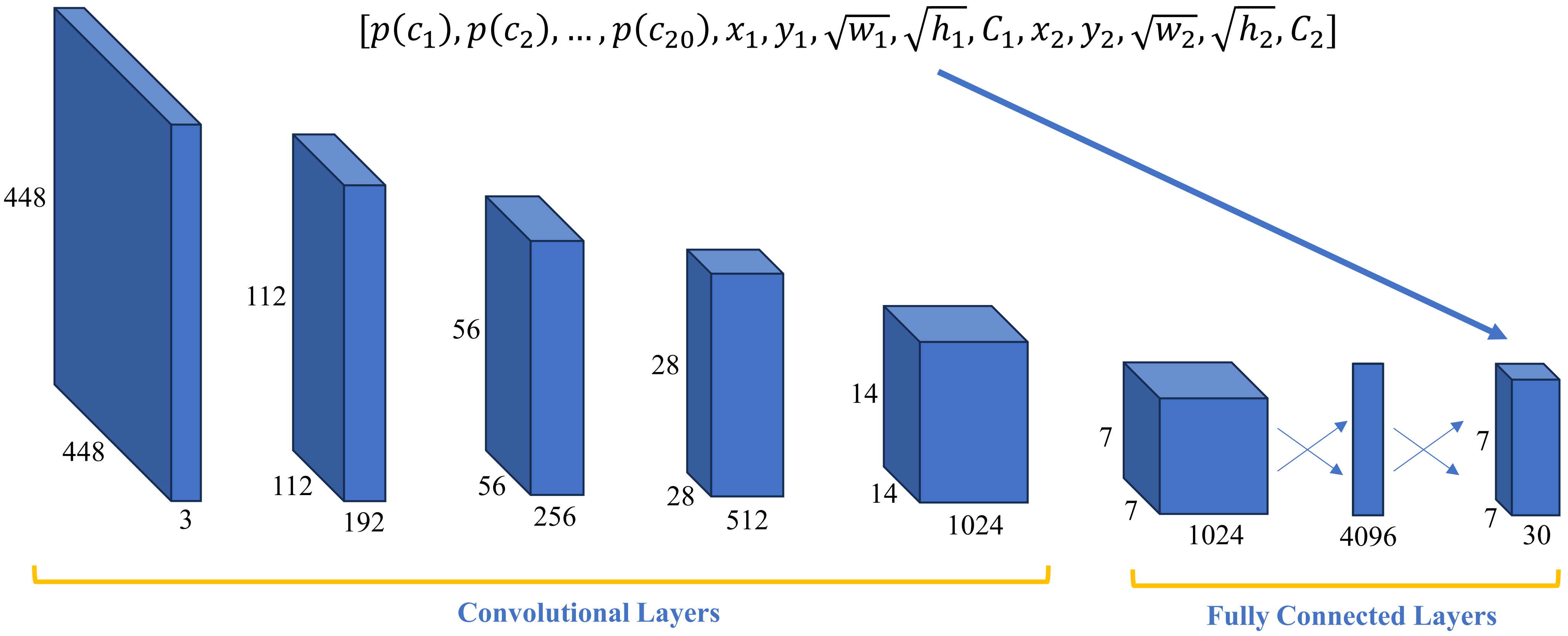
…

$[x_B, y_B, \sqrt{w_B}, \sqrt{h_B}, C_B]$

n is the number of object classes
B is the number of bounding boxes

Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does YOLO Work?

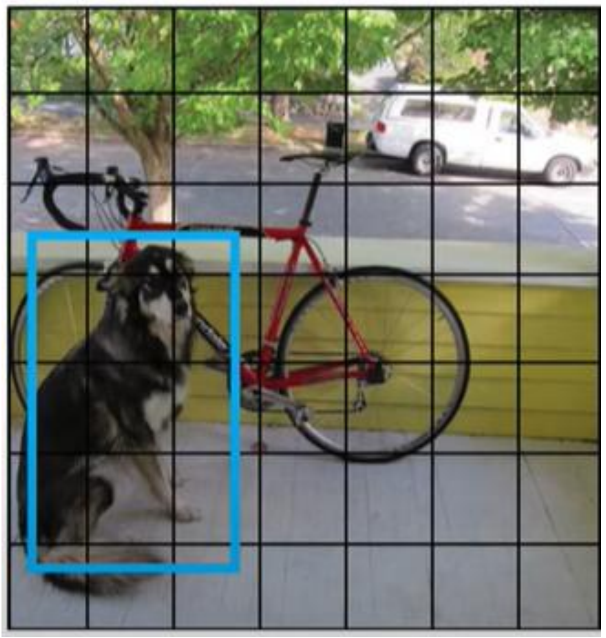


Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

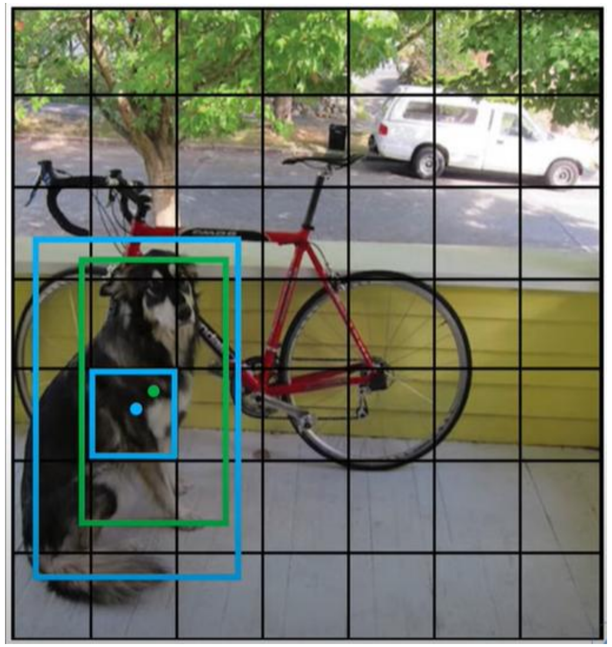
IPSA ÉCOLE D'INGÉNIEURS

How does YOLO Work?

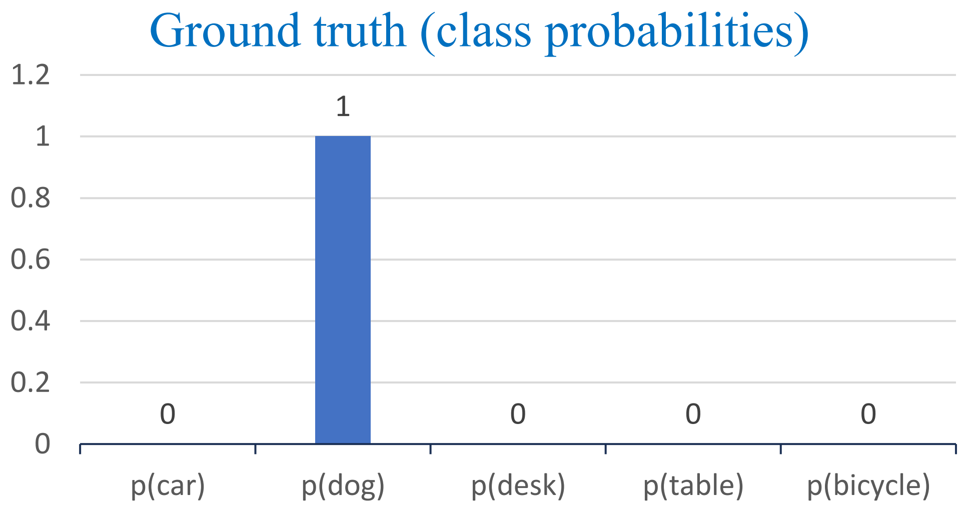
Ground truth (class probabilities)



Let’s predict this bounding box



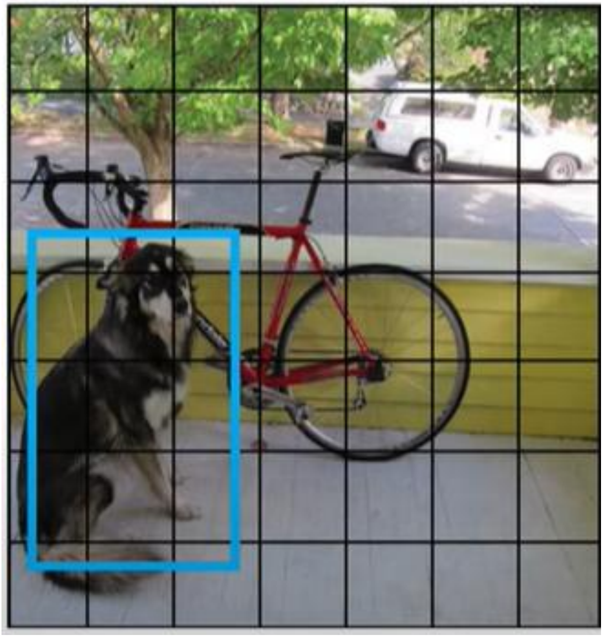
Green box is our prediction



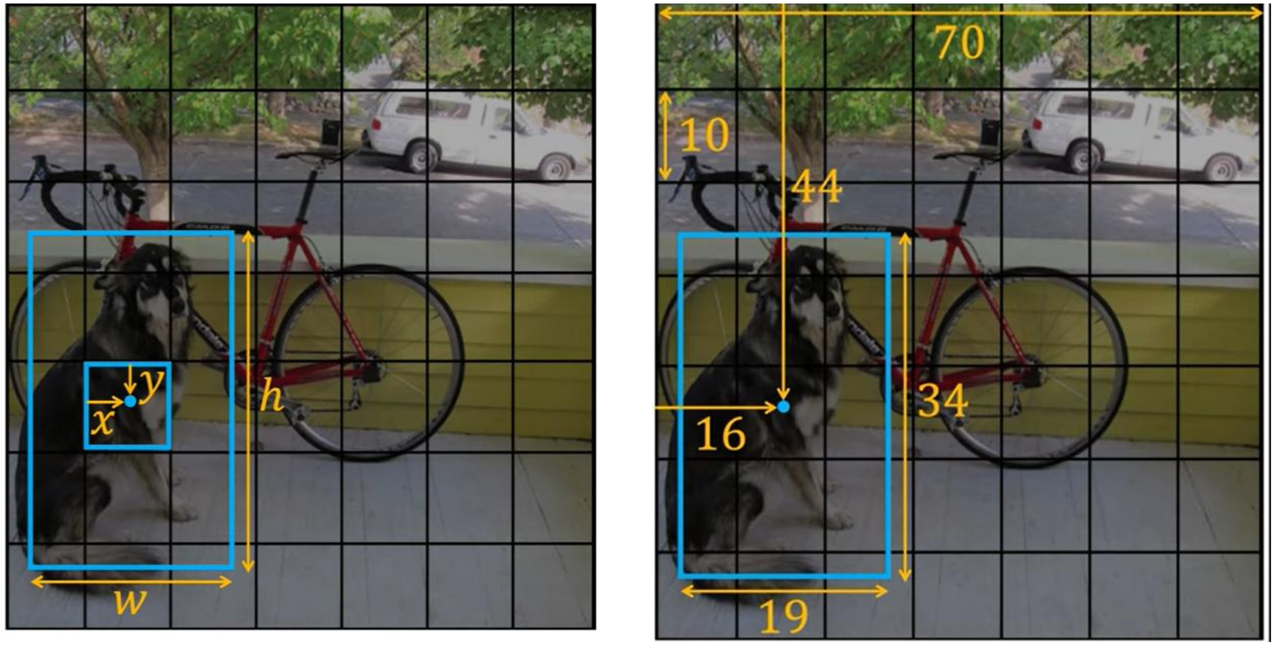
Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does YOLO Work?

Ground truth (box coordinates)



Let’s predict this bounding box



$$x = \frac{16\%10}{10} = 0.60$$

$$y = \frac{44\ \%10}{10} = 0.40$$

$$w = \frac{19}{70} = 0.27$$

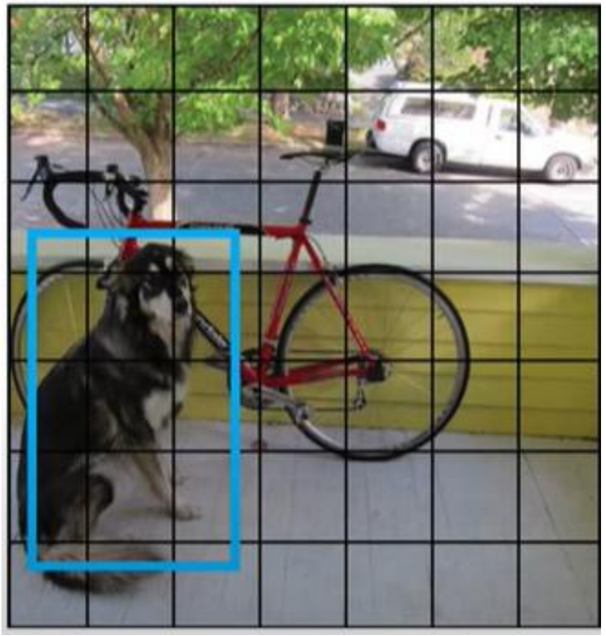
$$h = \frac{34}{70} = 0.49$$

Calculate the ground truth and predicted boxes by using the coordinates

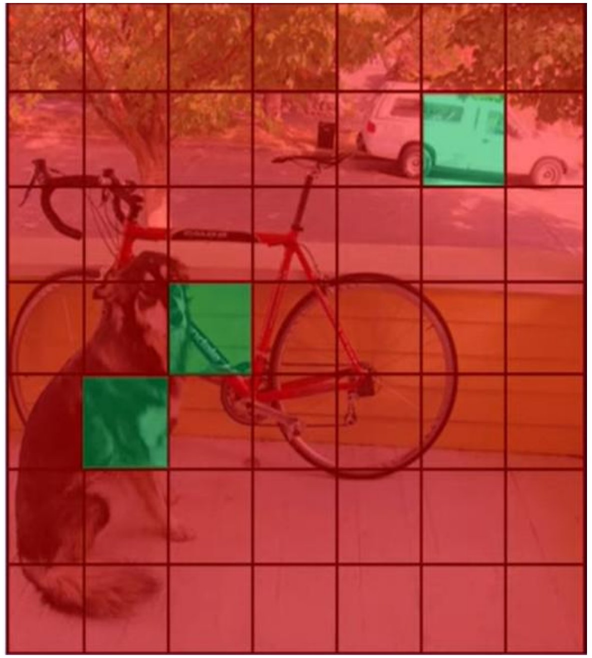
Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

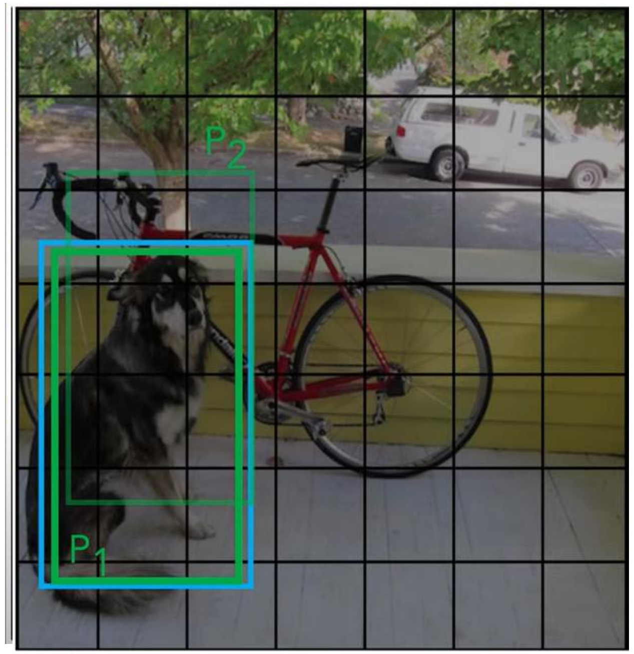
How does YOLO Work?

Ground truth (confidence)



Let's predict this bounding box





$= 0$

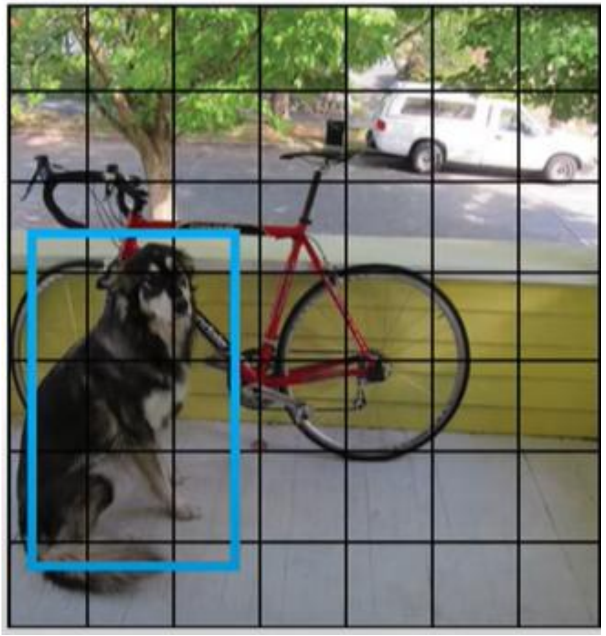
$= IoU\ (pred, true)$

$$C\ = IoU\ (pred, true)$$

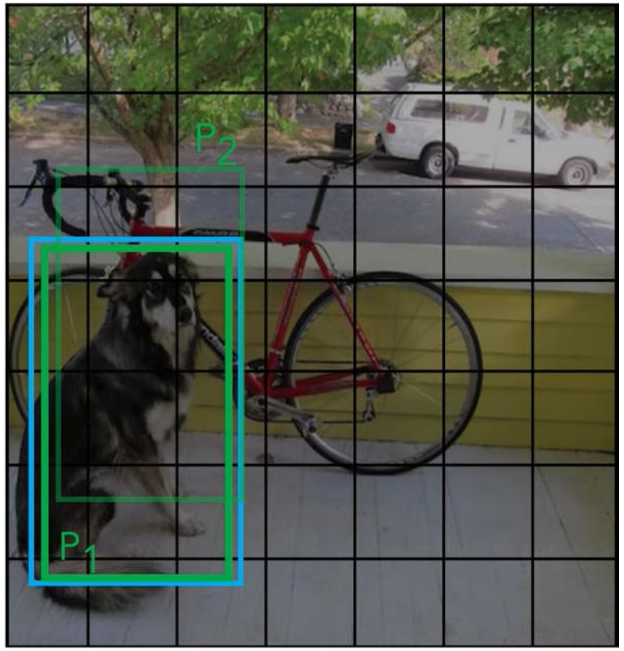
Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does YOLO Work?

Box selection (Inference)



Let's predict this bounding box



$If\ IoU\ (P_1, P_2\ ) > Threshold$:

$\mathrm{P} = \mathrm{argmax}(\mathrm{C}(P_1), C(P_2\ ))$

Source: https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s

How does YOLO Work?

For training part, Loss function

$$\boldsymbol{L} = \overbrace{\lambda_{coord}}^{\text{Helps to put more weight to box coordinates}} \times \overbrace{\sum_{i=1}^{S^2} 1_i^{obj} \times \left(\begin{array}{c} (\Delta x_i^* - \Delta \hat{x}_i)^2 + (\Delta y_i^* - \Delta \hat{y}_i)^2 + \\ \left(\sqrt{\Delta w_i^*} - \sqrt{\Delta \hat{w}_i}\right)^2 + \left(\sqrt{\Delta h_i^*} - \sqrt{\Delta h_i}\right)^2 \end{array} \right)}^{\text{Bounding box coordinate loss (Coordinate regression loss)}}$$

$$\underbrace{+ \sum_{i=1}^{S^2} 1_i^{obj} \times (c_i^* - \hat{c}_i)^2}_{\text{Confidence score loss}} \underbrace{+ \sum_{i=1}^{S^2} 1_i^{obj} \times \sum_{c=1}^{20} (p_{i,c} - \hat{p}_{i,c})^2}_{\text{Class probability loss}}$$

If there is an object, it is evaluated

$$+ \lambda_{no\_obj} \sum_{i=1}^{S^2} 1_i^{no\_obj} \times \sum_{j=1}^{B} (c_{i,j} - \hat{c}_{i,j})^2$$

If there is no object, it is evaluated

https://www.youtube.com/watch?v=zgbPj4lSc58&list=PL1u-h-YIOL0sZJsku-vq7cUGbqDEeDK0a

https://www.youtube.com/watch?v=svn9-xV7wjk&t=170s