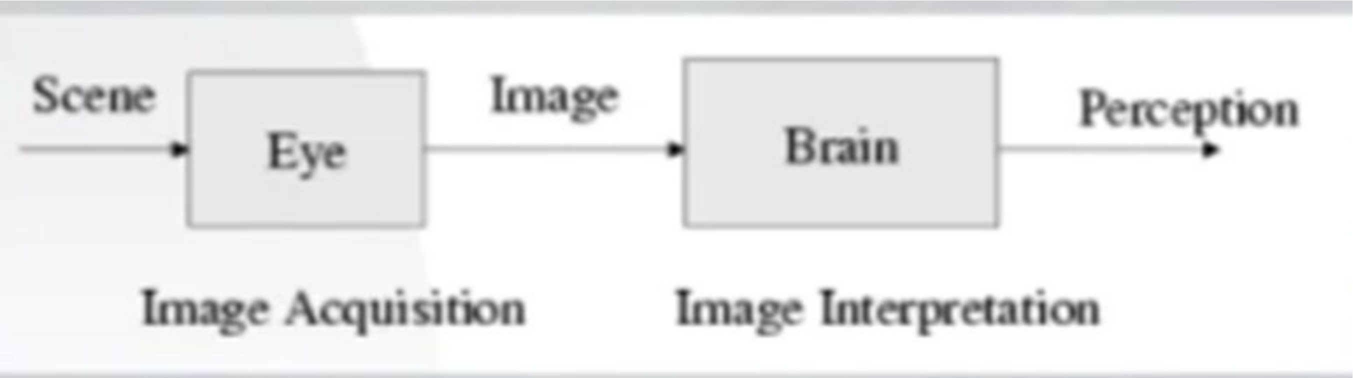
Object Recognition

Presented by: Group 6

1. Md. Fayjullah (19CSE008)
2. Hriday Krishna Bepari (19CSE018)
3. Sumya Habib Meem (19CSE024)
4. Anisur Rahman (19CSE046)
5. Obaydul Hasan Nayeem (19CSE014)

Human Vision

Vision is the process of discovering what is present in the world and where it is by looking.

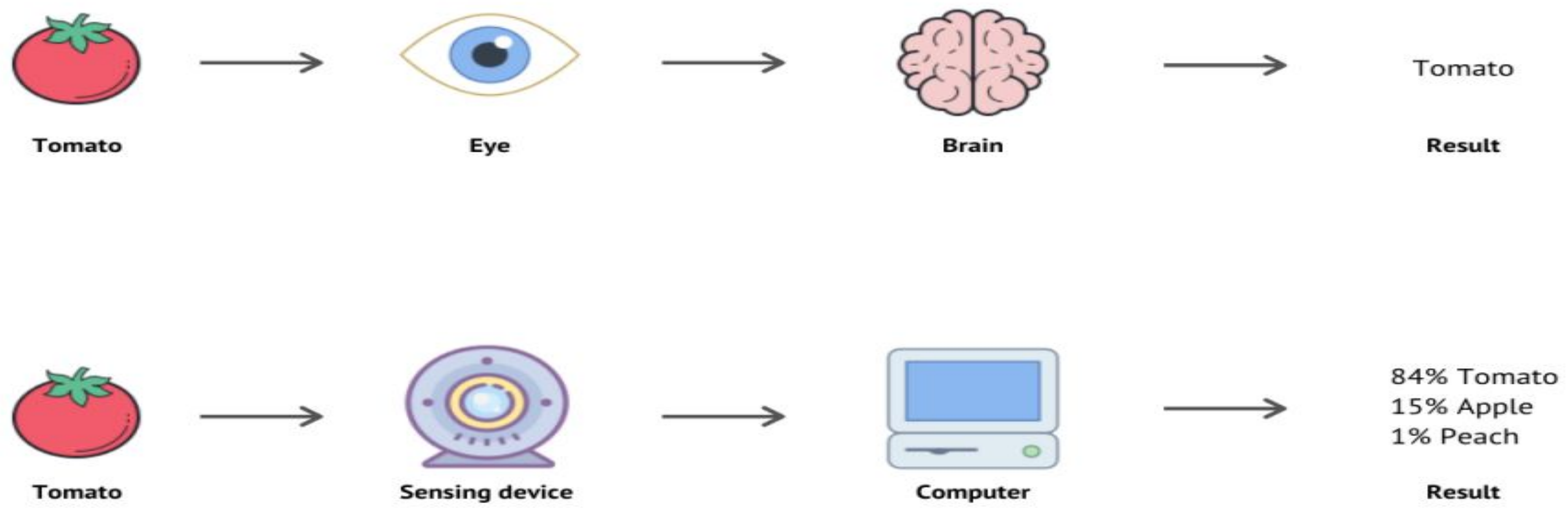


Computer Vision

Computer Vision is the study of analysis of pictures and videos in order to achieve results similar to those as by humans.

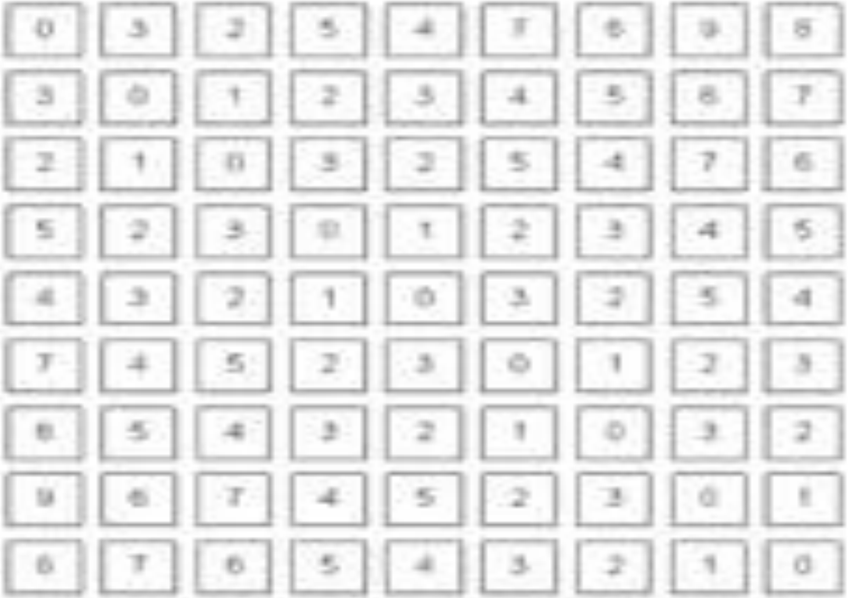
A field of computer science that focuses on enabling computers to identify and understand objects and people in images and videos.

Human Vision vs Computer Vision



Computer vision vs human vision

What we see



What a computer sees

What is Object Recognition?

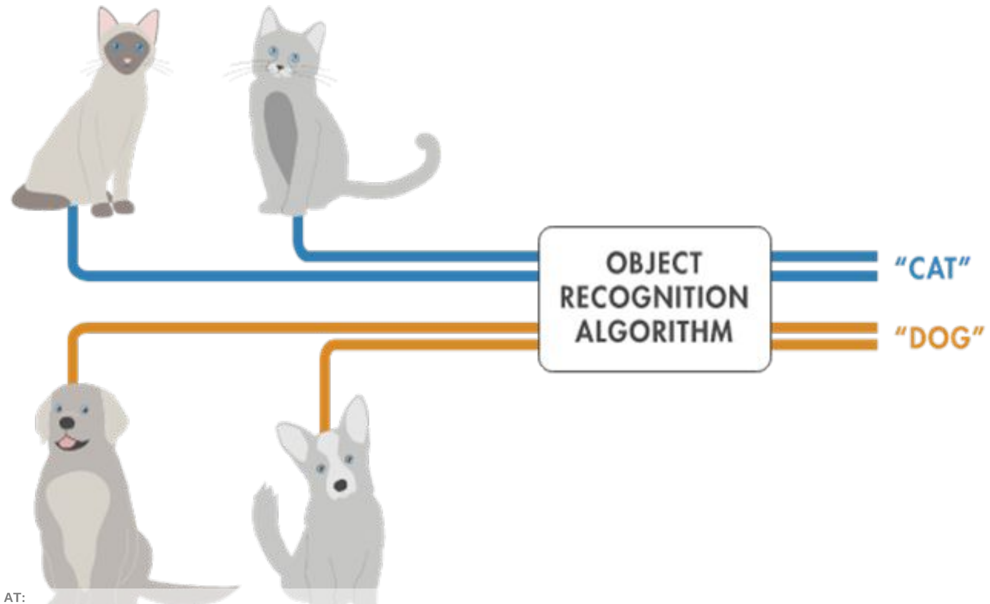
- Last step in image processing.
- It is the task of finding and identifying objects is in an image or video sequence.

What is Object Recognition? cont...

Like human understanding , it includes:

- **Detection -** of separate objects
- **Description -** of their geometry and positions in 3D
- **Classification -** as being one of a known class
- **Identification -** of the particular instance
- **Understanding -** of spatial relationships between objects

What is Object Recognition? cont...



What is Object Recognition? cont...

Object recognition consists of **recognizing, identifying, and locating objects** within a picture with a given degree of confidence. In this process, the four main tasks are:

- ❑ **Classification**
- ❑ **Tagging**
- ❑ **Detection**
- ❑ **Segmentation**

Object Recognition Use Cases

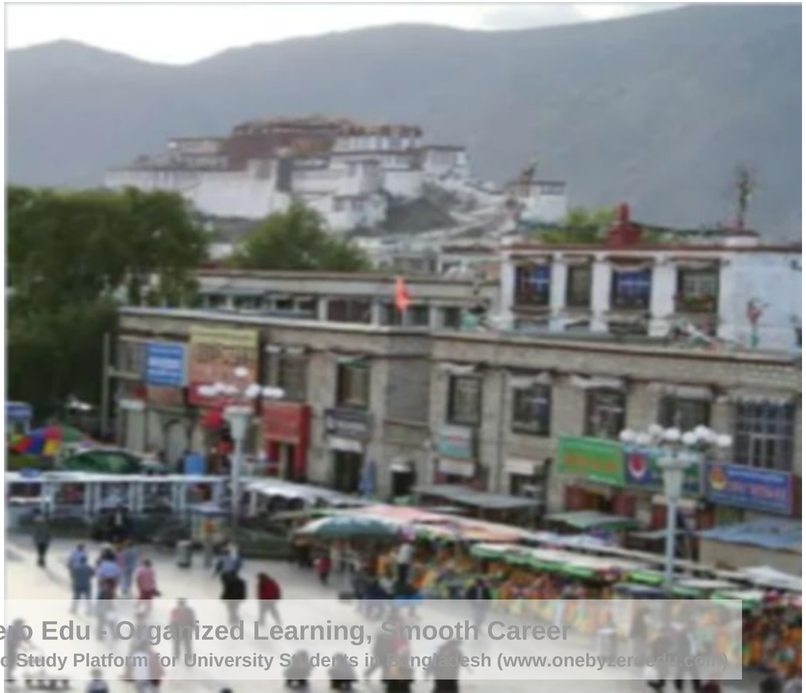
Automotive industry: This technology maintains self-driving cars' capability of recognizing objects around them and making safety-first decisions depending on the environment.

Security and surveillance: Private enterprises use object, face, and retina recognition technologies embedded in smart cameras for public safety.

Agriculture: Robotic tools empowered with computer vision can distinguish seedlings from weeds.

Retail: Smart cameras with a visual object recognition feature can detect customers and track their behavior in a store. Such insights are used to create better product and shelf placements.

So what does object recognition involve?



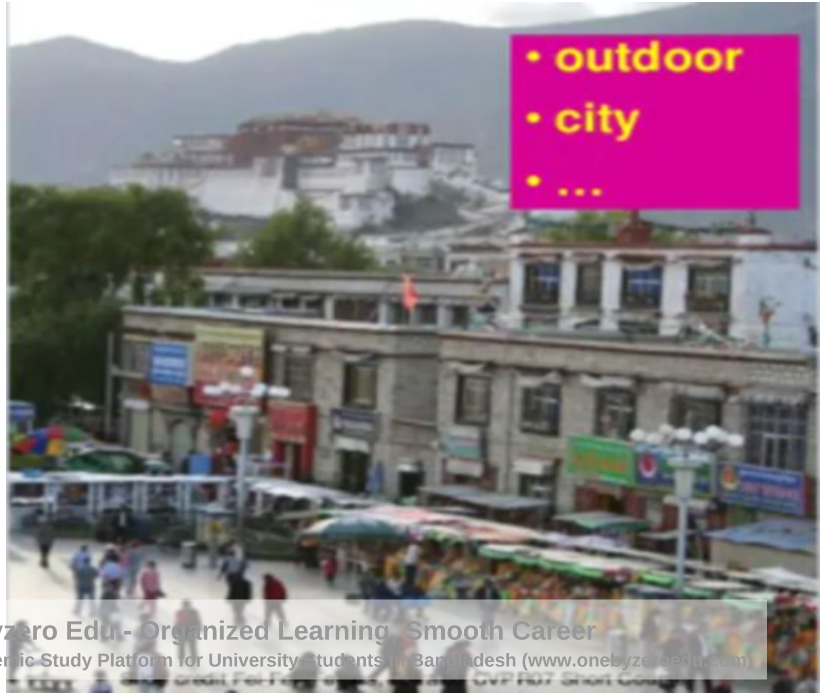
Detection: are there people?

Verification: Is that a lamp?

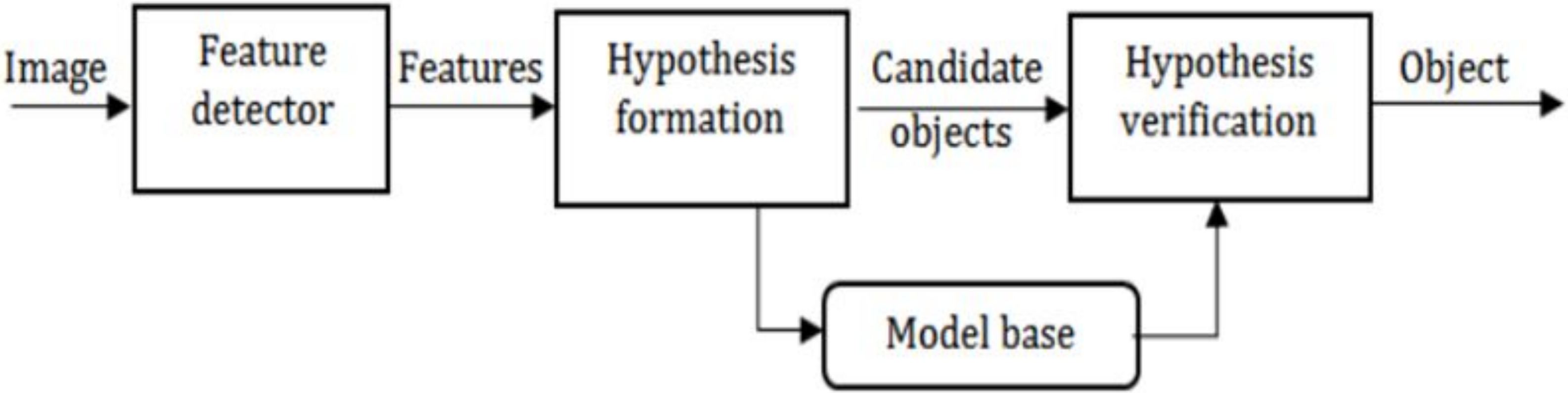
Identification: Is that Potala Place?

Object Classification

Scene and context classification / Understanding



Different components of Object Recognition



An object recognition system must have the following components to perform the task:

- ❖ Model database (also called modelbase)
- ❖ Feature detector
- ❖ Hypothesizer
- ❖ Hypothesis verifier

The central issues that should be considered in designing an object recognition system are:

Object or model representation:

How the objects or models to be recognized are represented in the system.

Example:

A system for recognizing fruits might use 3D models capturing their shapes, textures, and colors.

A system for recognizing handwritten digits might use template matching with pre-defined digit patterns.

Feature Extraction:

The process of identifying and extracting relevant features from raw data that can be used for recognizing objects.

- **Example:**

–In image-based object recognition, features like edges, corners, textures, and color histograms can be extracted.

Feature-Model Matching:

Comparing the extracted features with the stored models or representations of objects.

Example:

–For template matching, the extracted features are compared directly with the pre-defined templates.

–For more complex models, machine learning algorithms like convolutional neural networks (CNNs) can be used to learn the relationships between features and object classes.

Hypotheses Formation:

Generating possible hypotheses about the presence and identity of objects in the scene based on the matched features.

Example:

–If a system detects edges and corners in a specific arrangement, it might hypothesize the presence of a rectangular object.

Object Verification

Validating the hypotheses by checking additional evidence or refining the matching process to confirm or reject the presence of the object.

Example:

Additional constraints or contextual information might be used to refine hypotheses.

For instance, knowing the expected size of a fruit in a particular scene could help rule out unlikely candidates.

COMPLEXITY OF OBJECT RECOGNITION

A qualitative way to consider the complexity of the object recognition task would consider the following factors:

1. Scene Constancy:

High constancy: If the scene conditions are similar to the conditions in which the object models were created, the task is generally easier.

Low constancy: When the scene deviates significantly from the model conditions, it becomes more challenging to recognize objects due to variations in their appearance.

2. Image-Model Spaces:

Similar spaces: If the image and model spaces share similar characteristics (e.g., both are 2D images or 3D models), the recognition task is generally easier.

Dissimilar spaces: When the image and model spaces differ significantly (e.g., 2D image vs. 3D model), the task becomes more complex due to the need for additional processing and transformation steps.

3. Number of Objects in the Model Database:

Smaller database: With a smaller number of objects in the model database, the system has fewer possibilities to consider, making the recognition task potentially easier.

Larger database: A larger database increases the complexity as the system needs to differentiate between a wider range of objects, potentially leading to confusion and misidentification.

4. Number of Objects in an Image and Possibility of Occlusion:

Single object, no occlusion: Recognizing a single object in an image without occlusion is generally the simplest scenario.

Multiple objects, potential occlusion: The presence of multiple objects and the possibility of occlusion significantly increase the complexity. The system needs to handle overlapping objects, partial visibility, and potential confusion between similar objects.

QUIZ:

1. **What is the primary purpose of object or model representation in an object recognition system?**

   A) To capture and store the essential features and characteristics of objects.

   B) To detect edges in images.

   C) To generate possible hypotheses about object presence.

   D) To validate the detected objects.

2. **How does occlusion affect the complexity of object recognition?**

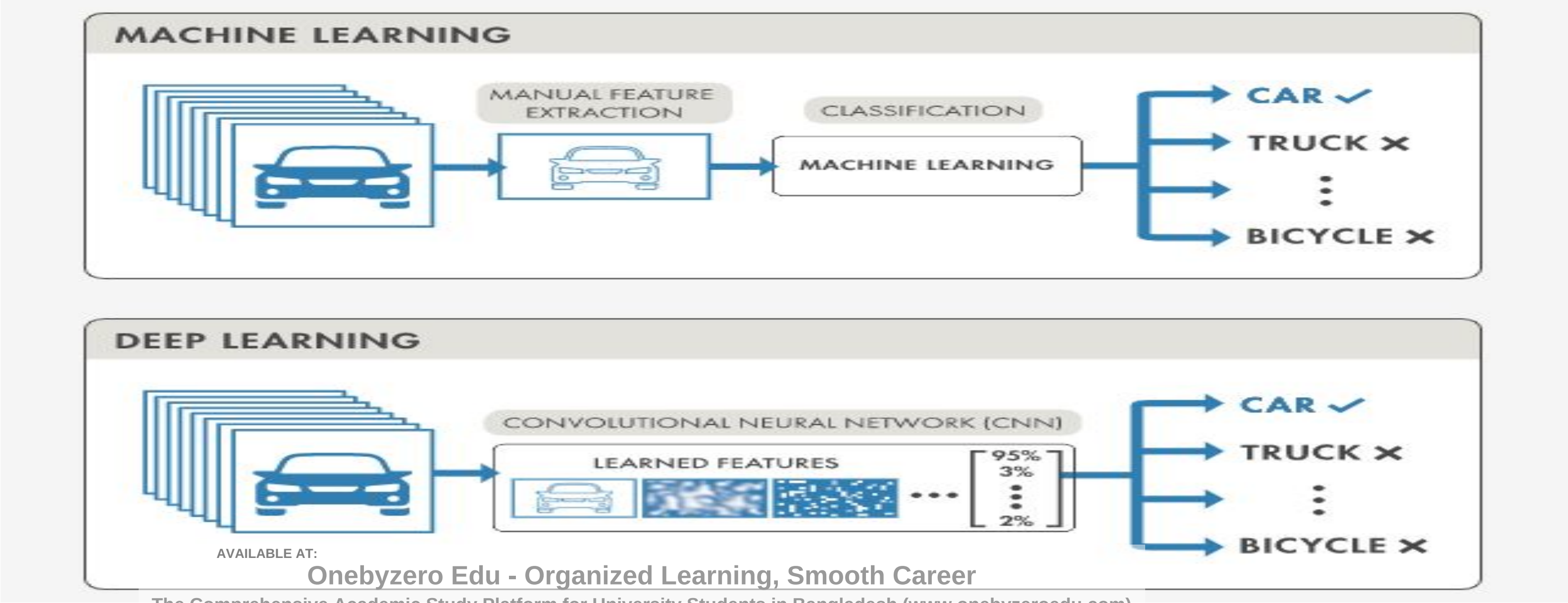
   A) It simplifies the feature extraction process.

   B) It makes it harder to identify and match features accurately.

   C) It eliminates the need for hypotheses formation.

   D) It has no significant impact.

How Object Recognition Works



How Object Recognition Works

Object detection can be performed using either traditional -

(1) Machine Learning

(2) Deep Learning Networks

Machine Learning

Machine Learning means computers learning from data using algorithms to perform a task without being explicitly programmed.Machine learning techniques are popular for object recognition and offer different approaches than deep learning. Common examples of machine learning techniques is SVM machine learning model

Deep Learning Networks

Deep learning techniques have become a popular method for doing object recognition. Deep learning models such as convolutional neural networks, or CNNs, are used to automatically learn an object's inherent features in order to identify that object. For example, a CNN can learn to identify differences between cats and dogs by analyzing thousands of training images and learning the features that make cats and dogs different.

Two approaches to performing object recognition using deep learning

Training a model from scratch: To train a deep network from scratch, you gather a very large labeled dataset and design a network architecture that will learn the features and build the model. The results can be impressive, but this approach requires a large amount of training data.

Using a pre trained deep learning model: Most deep learning applications use the transfer learning approach, a process that involves fine-tuning a pretrained model.This method is less time-consuming and can provide a faster outcome because the model has already been trained on thousands or millions of images.

Quiz

Which of the statements below is false?

1. Being able to try out ideas quickly allows deep learning engineers to iterate more quickly.
2. Faster computation can help speed up how long a team takes to iterate to a good idea.
3. It is faster to train on a big dataset than a small dataset.
4. Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware).

Classification vs Localization

Classification

[{CAT}]

Classification + Localization

[{CAT,(x,y,h,w)}]

Object Detection

[{DOG,(x,y,h,w)},
{DOG,(x,y,h,w)},
{CAT,(x,y,h,w)}]

Image Classification

In Image classification, it takes an image as an input and outputs the classification label of that image with some metric (probability, loss, accuracy, etc). For Example: An image of a cat can be classified as a class label “cat” or an image of Dog can be classified as a class label “dog” with some probability.

Object Localization

It locates the presence of an object in the image and represents it with a bounding box. It takes an image as input and outputs the location of the bounding box in the form of (position, height, and width).

Object Detection

Object Detection algorithms act as a combination of image classification and object localization. It takes an image as input and produces one or more bounding boxes with the class label attached to each bounding box. These algorithms are capable enough to deal with multi-class classification and localization as well as to deal with the objects with multiple occurrences.

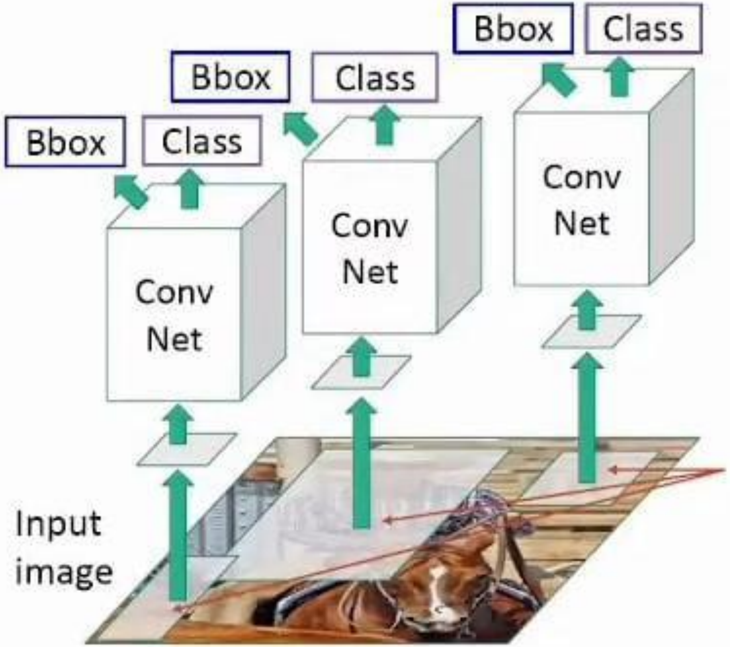
Most popular Algorithm

1. Region-Based Convolutional Neural Network (R-CNN)
2. Fast R-CNN
3. Faster R-CNN
4. You Only Look Once(YOLO)

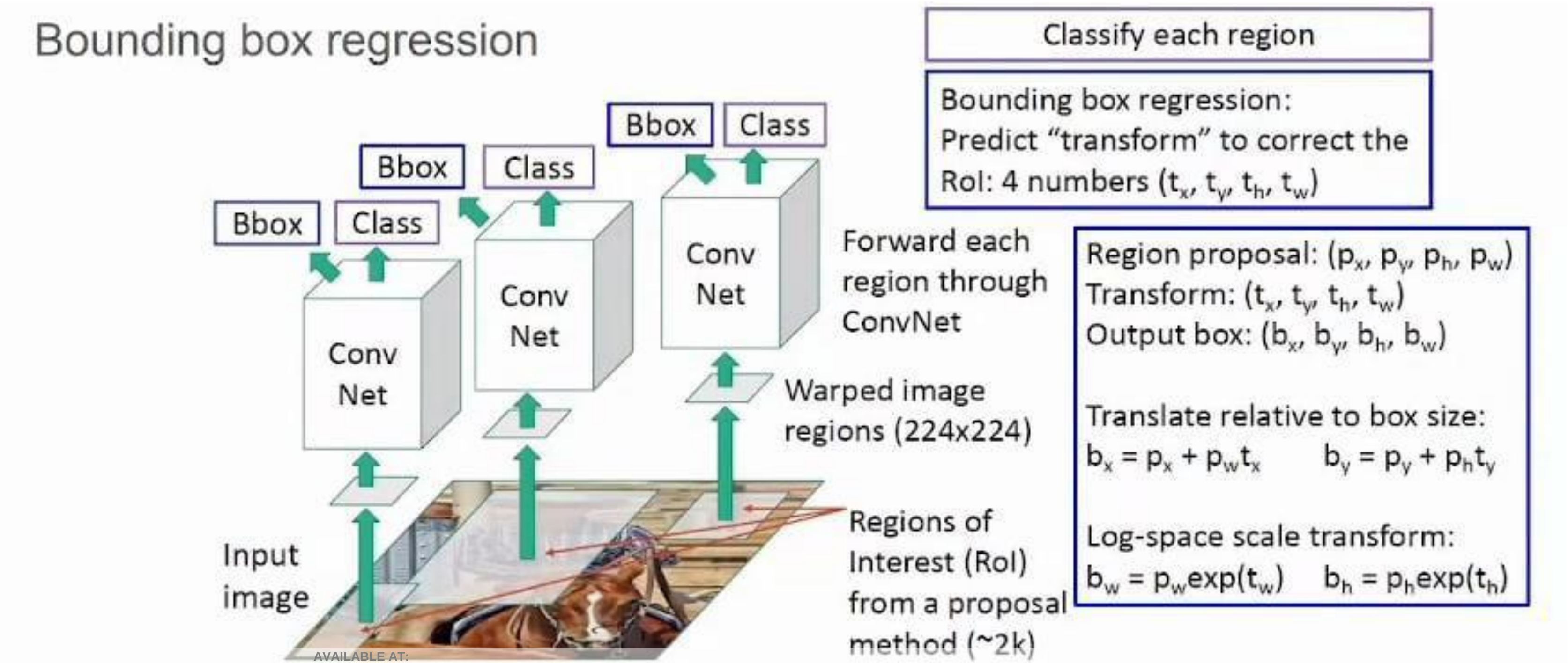
Region Based CNN (R-CNN)

Inference time: for an RGB input image

1. Run region proposal method to compute ~2000 region proposals
2. Resize each region to 224x224 and run independently through CNN to predict **class scores** and **bbox transform**
3. Use scores to **filter** a subset of region proposals to output
4. *Compare* with ground-truth boxes



Region Based CNN (R-CNN)

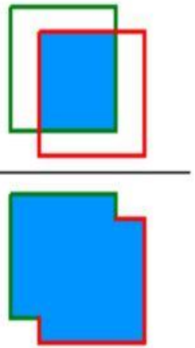


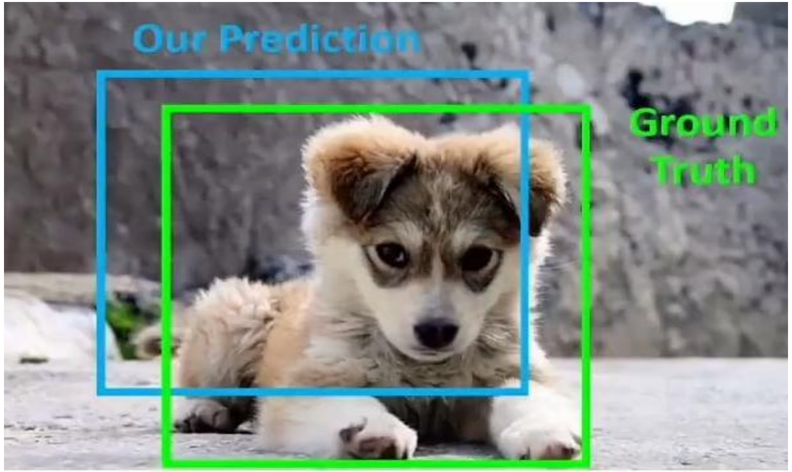
How to evaluate prediction ?

- ➢ Intersection Over Union (IOU)
- ➢ Non Maximum Suppression (NMS)
- ➢ Average Precision (AP)

Intersection Over Union (IOU)

$$IOU = \frac{\text{area of overlap}}{\text{area of union}} =$$





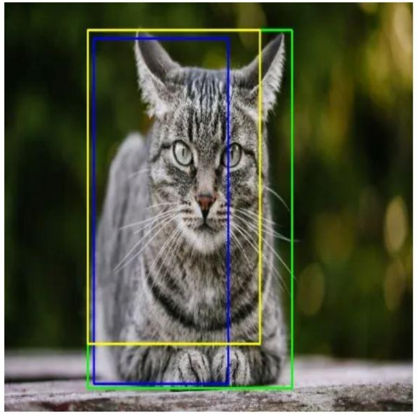
Intersection Over Union (IOU)

Intersection over Union (IoU) is a measure that shows how well the prediction bounding box aligns with the ground truth box. It's one of the main metrics for evaluating the accuracy of algorithms and helps distinguish between "correct detection" and "incorrect detection".

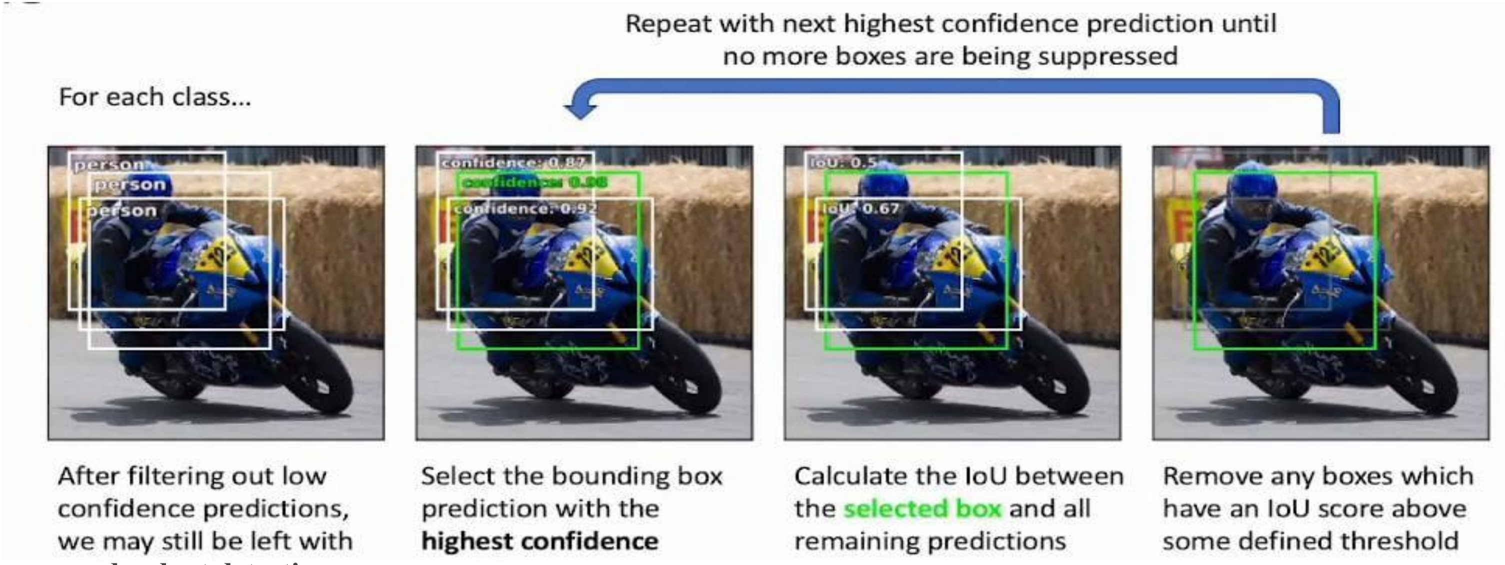
The IoU value ranges from 0 to 1, where 0 indicates no overlap between the boxes, and 1 represents complete overlap or identical regions. An Intersection over Union score > ***0.5*** is normally considered a **"good"** prediction.

Non Maximum Suppression (NMS)

Non max suppression is a technique used mainly in object detection that aims at selecting the best bounding box out of a set of overlapping boxes. In the following image, the aim of non max suppression would be to remove the yellow, and blue boxes, so that we are left with only the green box.



Non Maximum Suppression (NMS)



redundant detections

Non Maximum Suppression (NMS)

Algorithm Step:

1. Define a value for Confidence_Threshold, and IOU_Threshold.
2. Sort the bounding boxes in a descending order of confidence.
3. Remove boxes that have a confidence < Confidence_Threshold
4. Loop over all the remaining boxes, starting first with the box that has highest confidence.
5. Calculate the IOU of the current box, with every remaining box that belongs to the same class.
6. If the IOU of the 2 boxes > IOU_Threshold, remove the box with a lower confidence from our list of boxes.
7. Repeat this operation until we have gone through all the boxes in the list.

Quiz

Q: If you have two bounding boxes with an IoU of 0.6 and a confidence score of 0.8 and 0.9 respectively, and you apply NMS with an IoU threshold of 0.5, what happens?

a) Both boxes are kept

b) The box with the confidence score of 0.8 is suppressed

c) The box with the confidence score of 0.9 is suppressed

d) Both boxes are suppressed

NMS Limitation

Problem: NMS may eliminate "good" boxes when objects are highly overlapping .

Concept: TP,FP,FN,P,R

True Positive(TP): Detection with IoU>= threshold.

False Positive(FP): An incorrect detection of a nonexistent object or a misplaced detection of an existing object. Detection with IoU < threshold.

False Negative(FN): An undetected ground-truth bounding box.

True Negative (TN): Correctly identifying that no object is present in a location where there is no ground-truth bounding box.

Concept: Precision & Recall

Precision : It is the percentage of correct positive predictions and is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{all detections}}$$

Recall: It is the percentage of true positive detected among all relevant ground truths and is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{all ground truths}}$$

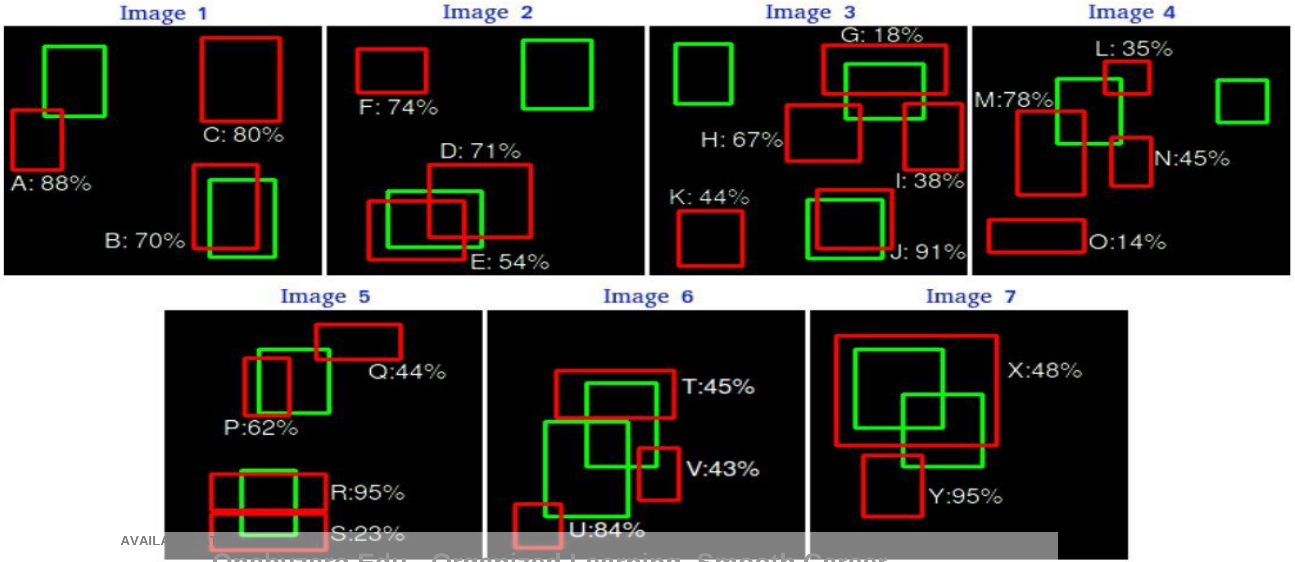
Average Precision (AP)

For each detection of a single category(highest score to lowest score)

1. If it matches some Ground-Truth box with Iou > 0.5 mark it as true positive
2. Otherwise mark it as Negative (FP) i.e Iou < 0.5
3. Plot a point on PR Curve
4. Average Precision(AP)= area under PR curve

Average Precision (AP)

Example: There are 7 images with 15 ground truth objects represented by the green bounding boxes and 24 detected objects represented by the red bounding boxes. Each detected object has a confidence level and is identified by a letter (A,B,...,Y).



Average Precision (AP)

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	[illegible]	48%	[illegible]
Image 7	[illegible]	[illegible]	FP

The following table shows the bounding boxes with their corresponding confidences. The last column identifies the detections as TP or FP. In this example a TP is considered if IoU 30%, otherwise it is a FP.

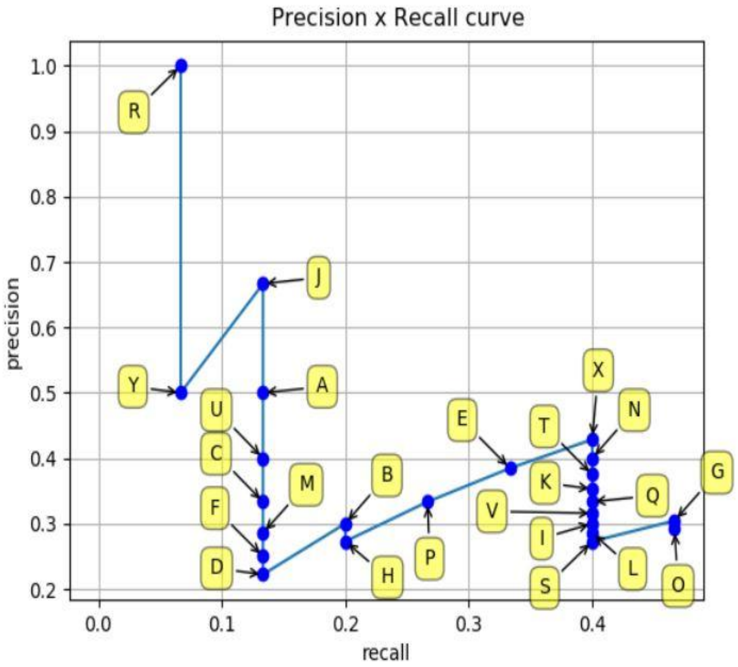
Average Precision (AP)

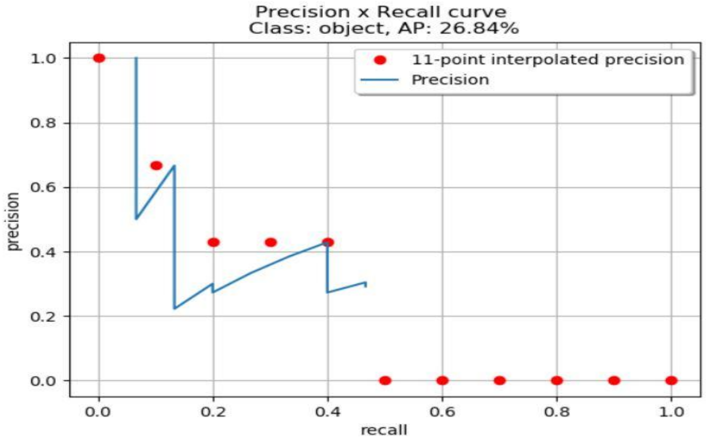
- First we need to order the detections by their confidences, then we calculate the precision and recall for each accumulated detection.

- Plotting the precision and recall values we have the following Precision vs Recall curve.

Images	Detections	Confidences	TP	FP	Acc TP	Acc FP	Precision	Recall
Image 5	R	95%	1	0	1	0	1	0.0666
Image 7	Y	95%	0	1	1	1	0.5	0.0666
Image 3	J	91%	1	0	2	1	0.6666	0.1333
Image 1	A	88%	0	1	2	2	0.5	0.1333
Image 6	U	84%	0	1	2	3	0.4	0.1333
Image 1	C	80%	0	1	2	4	0.3333	0.1333
Image 4	M	78%	0	1	2	5	0.2857	0.1333
Image 2	F	74%	0	1	2	6	0.25	0.1333
Image 2	D	71%	0	1	2	7	0.2222	0.1333
Image 1	B	70%	1	0	3	7	0.3	0.2
Image 3	H	67%	0	1	3	8	0.2727	0.2
Image 5	P	62%	1	0	4	8	0.3333	0.2666
Image 2	E	54%	1	0	5	8	0.3846	0.3333
Image 7	X	48%	1	0	6	8	0.4285	0.4
Image 4	N	45%	0	1	6	9	0.4	0.4
Image 6	T	45%	0	1	6	10	0.375	0.4
Image 3	K	44%	0	1	6	11	0.3529	0.4
Image 5	Q	44%	0	1	6	12	0.3333	0.4
Image 6	V	43%	0	1	6	13	0.3157	0.4
Image 3	I	38%	0	1	6	14	0.3	0.4
Image 4	L	35%	0	1	6	15	0.2857	0.4
Image 5	S	23%	0	1	6	16	0.2727	0.4
[illegible]	[illegible]	18%	1	0	7	16	0.3043	0.4666
[illegible]	[illegible]	[illegible]	[illegible]	1	7	17	0.2916	0.4666

Average Precision (AP)





By applying the 11-point interpolation, we have:

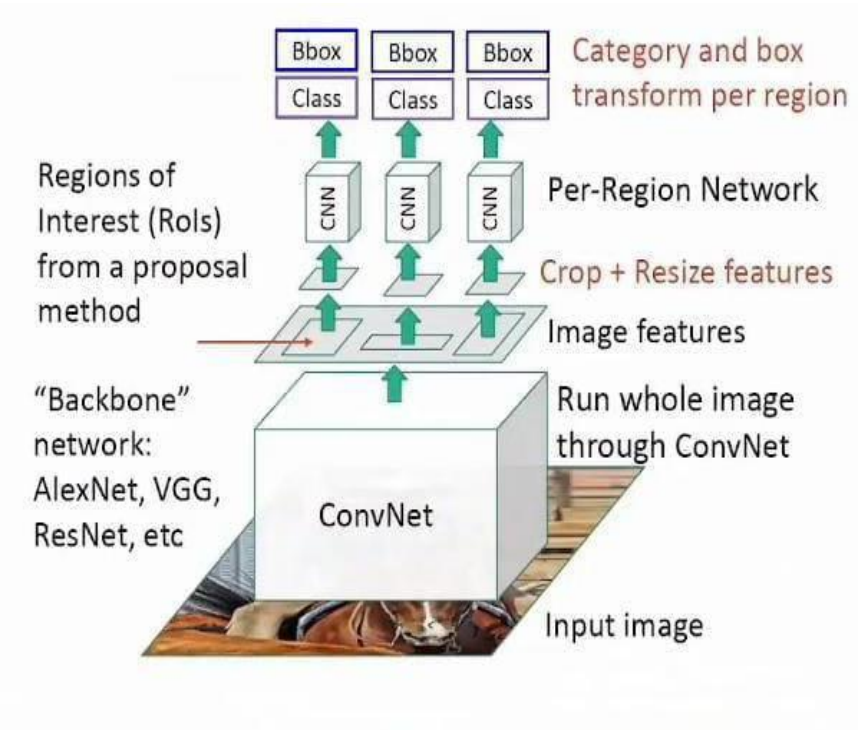
$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} \rho_{\text{interp}(r)}$$

$$AP = \frac{1}{11}(1 + 0.6666 + 0.4285 + 0.4285 + 0.4285 + 0 + 0 + 0 + 0 + 0 + 0)$$

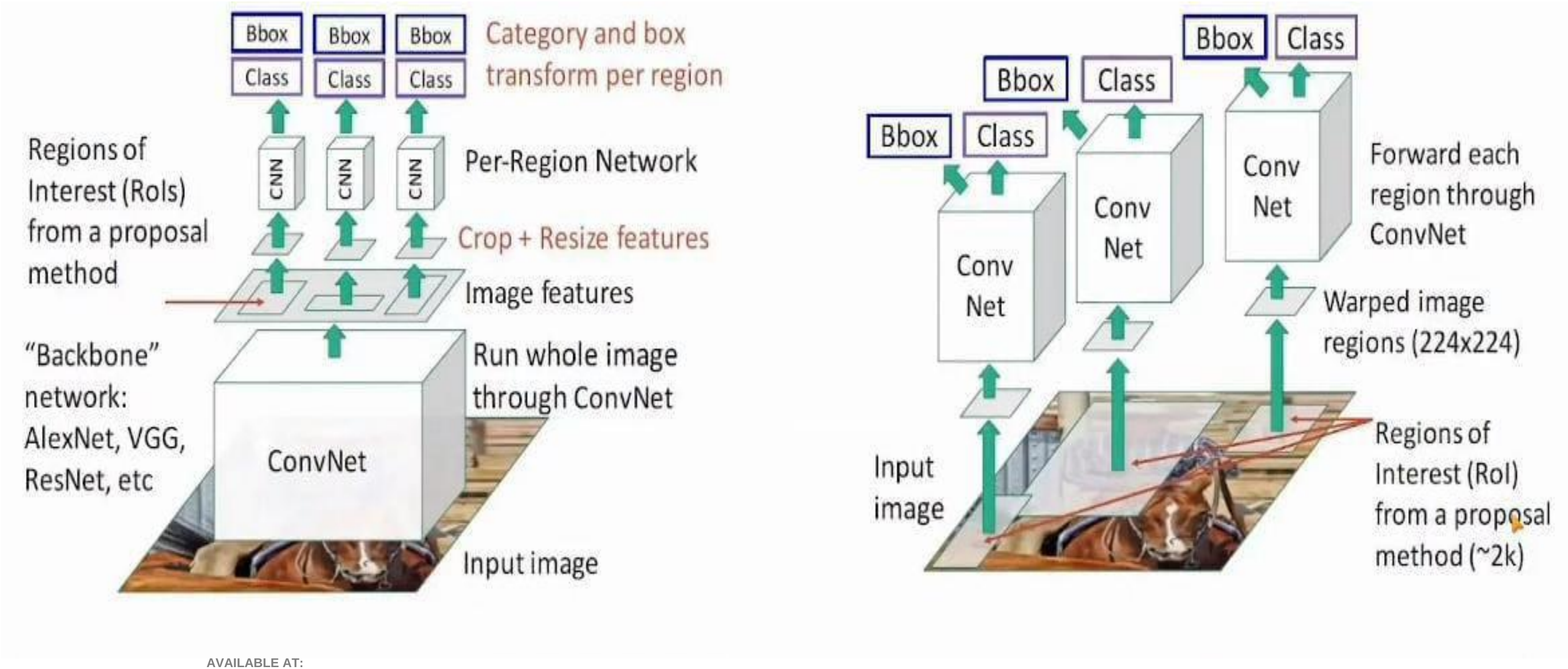
$$AP = 26.84\%$$

Fast R CNN

- An input image and multiple regions of interest(RoIs) are input into a fully cnn.
- Each roI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCS).
- The network has output vector per-class bounding-box.

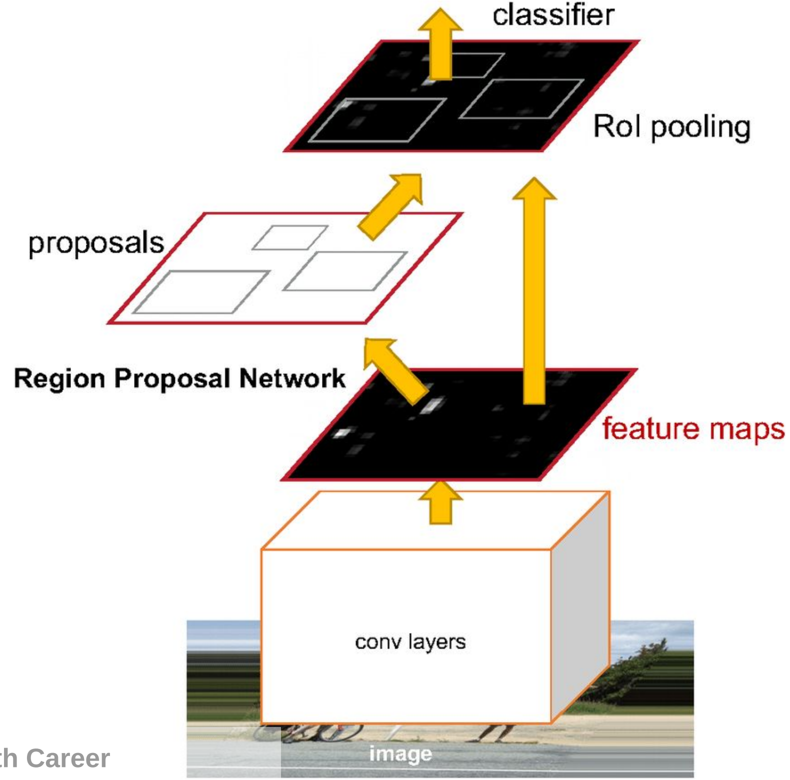


Fast R-CNN Vs R-CNN



Faster R-CNN

1. Regions of interest are generated using the region proposal network (RPN). To generate RoIs, the RPN uses convolutional layers.
2. The second part of Faster R-CNN is classification. It outputs the final bounding boxes and accepts two inputs—the list of RoIs from the previous step (RPN), and a feature volume computed from the input image.

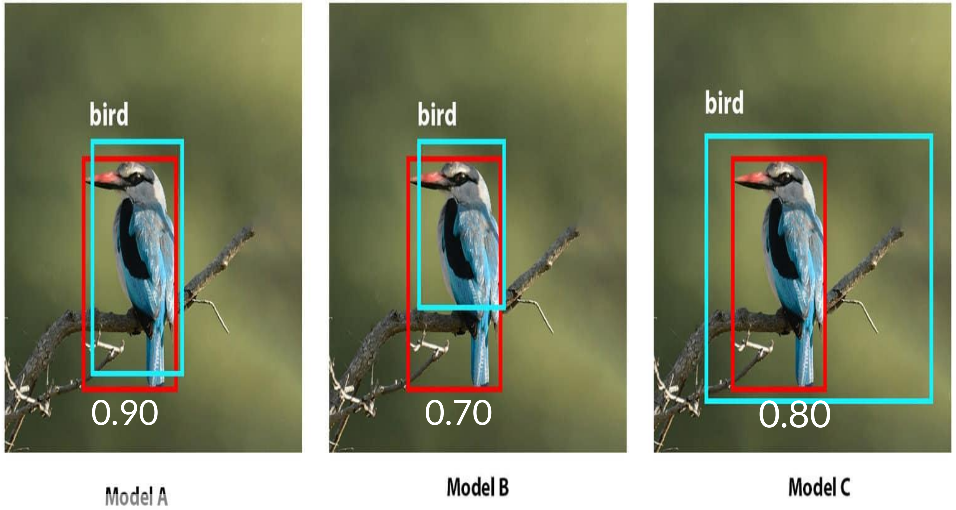


Code

https://colab.research.google.com/drive/1UI5iuIW0rSoIvh3KeyxpDfHjhSYcR3yZ?usp=sharing

Assignment

Calculate the Average Precision (AP) for the given two image. Each detection includes a confidence score and whether it is a true positive (TP) or a false positive (FP).



Home Work

1.What kind of output does an object recognition algorithm provide, and provide one real-world application of object recognition? [5]

2.In evaluation metrics, IoU, NMS and AP, which one is better for assessing the performance of models and why? [4]

3. Explain the differences between R-CNN, Fast R-CNN, and Faster R-CNN in terms of their architecture, efficiency, and performance. [5]

4. Explain the limitation of Non Maximum Suppression (NMS). [3]

Object Recognition Using YOLO Algorithm in Image Processing

What is YOLO algorithm?

YOLO: You Only Look Once

The YOLO algorithm divides the input image into a grid of cells, and for each cell, it predicts the probability of the presence of an object and the bounding box coordinates of the object.

Why YOLO Algorithm is used in Object Recognition?

YOLO can be applied to:

- Image File
- Webcam Feed
- Video File

- Extremely fast (45 frames per second).
- Realtime performance
- High accuracy
- Applications across various domains

Pre-trained YOLOv3 can detect 80 objects, such as:

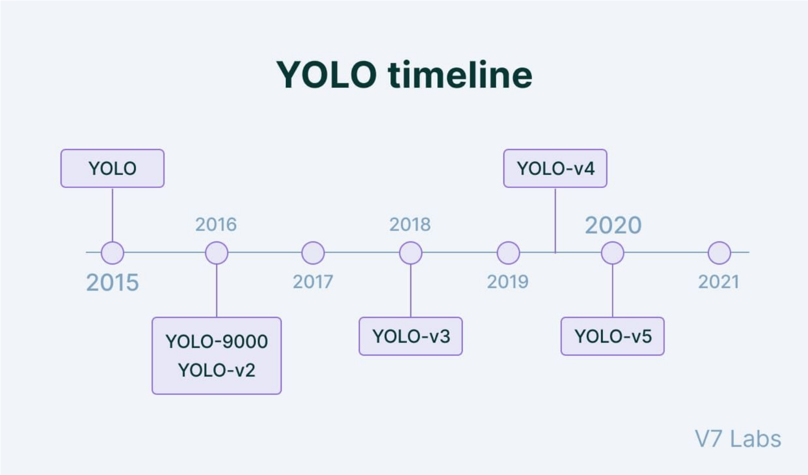
History & Development

Initial Development:

Created by Joseph Redmon and Ali Farhadi in 2016

Evolution:

YOLOv1 to YOLOv8: Key improvements and milestones



Comparison of YOLO with Other Algorithms

Algorithm	VOC 2007 AP	Picasso AP	People-Art AP
YOLO	59.2	53.3	45
R-CNN	54.2	10.4	26
DPM	43.2	37.8	32

https://pjreddie.com/media/files/papers/yolo.pdf

How YOLO Algorithm is used in Object Recognition

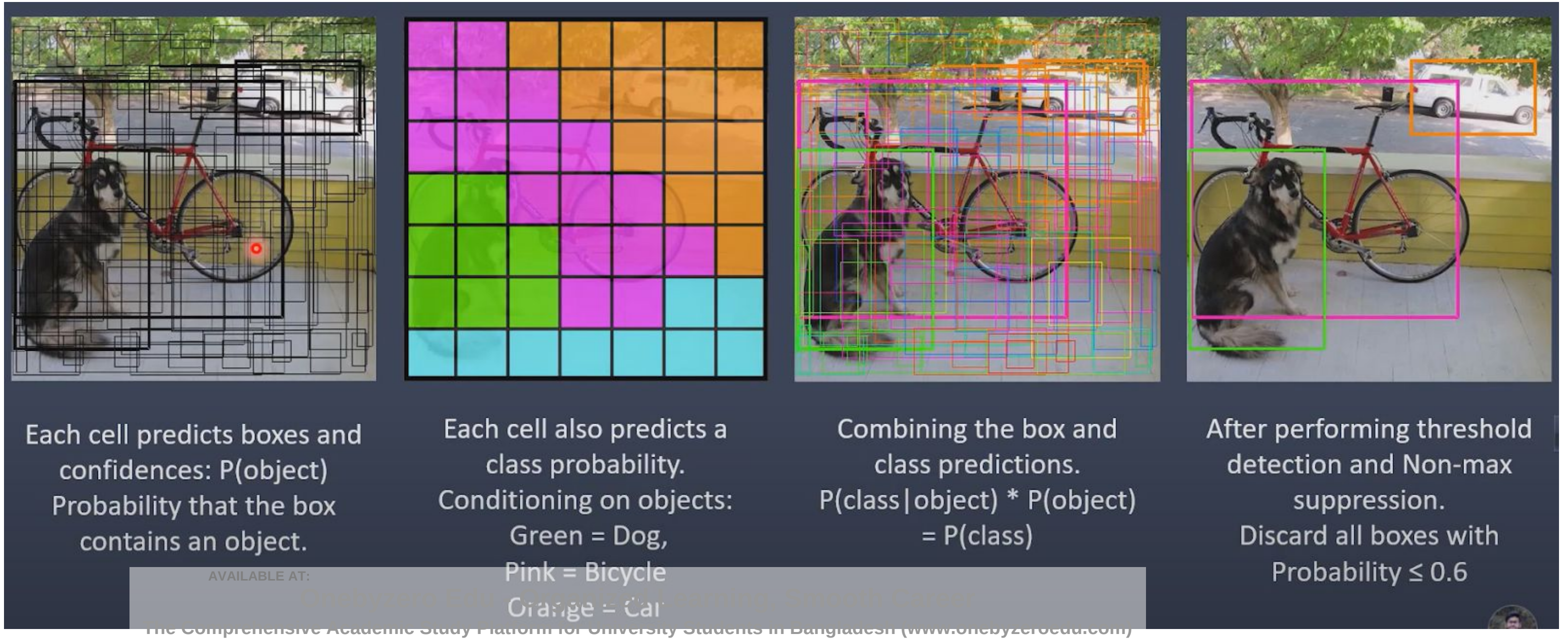
Image is divided into s*s grid

Each cell predicts B boxes (x,y,w,h) and confidence of each box: P(object)

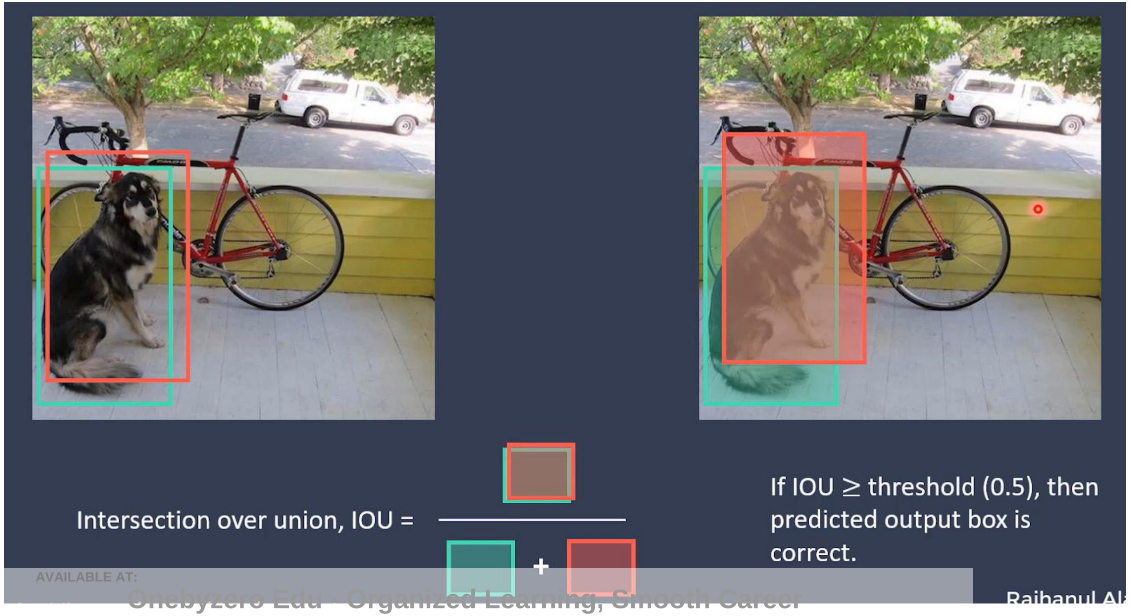
Predicted box center is shown in black which is in the blue colored cell.

Anchor box specifying the object in a grid cell.

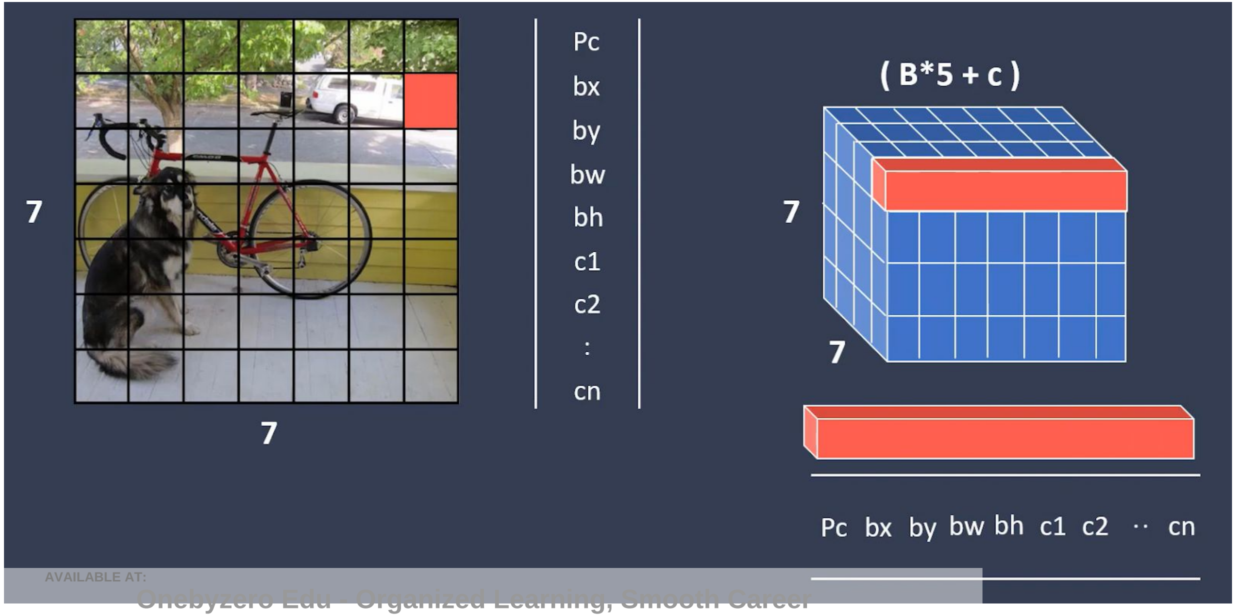
How YOLO Algorithm is used in Object Recognition



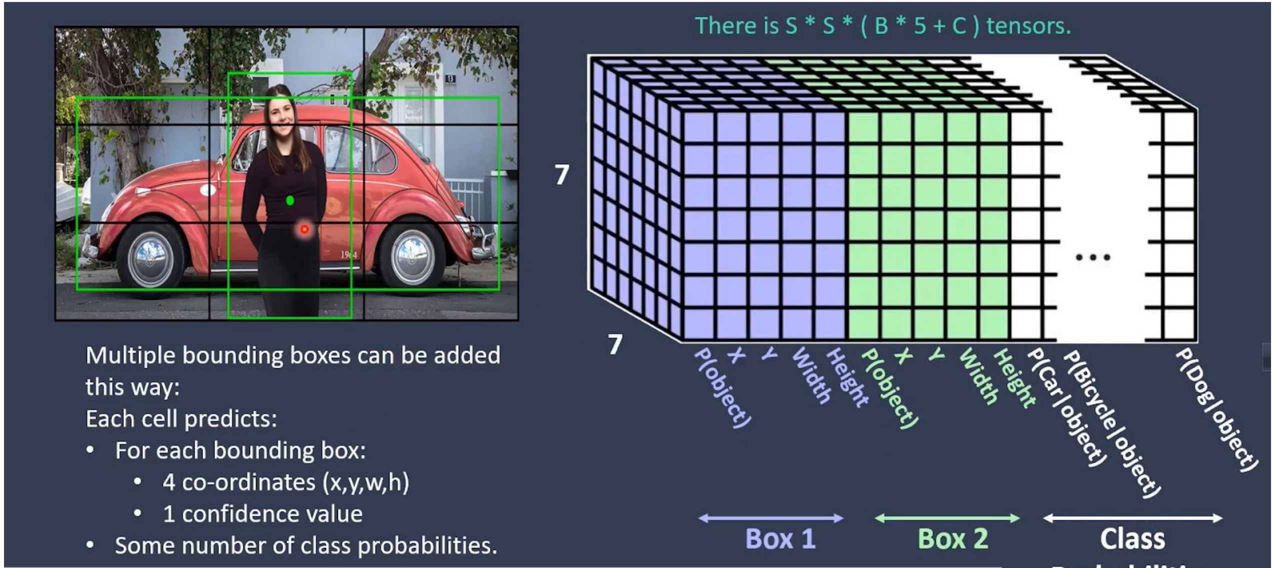
How YOLO Algorithm is used in Object Recognition



How YOLO Algorithm is used in Object Recognition



How YOLO Algorithm is used in Object Recognition



AVAILABLE AT:
https://pjreddie.com/media/files/papers/yolo.pdf

Thank You