Artificial Intelligence

Computer Vision – Object Recognition

Dr. Sarwan Singh

Scientist – D, NIELIT Chandigarh

AI-Roadmap to Computer Vision

1. Python & Statistics : Basics of computer vision
2. Solving Image Classification : Basic Machine Learning Image pre-processing Techniques
3. Introduction to Keras & Neural Networks: Fundamentals of Neural Networks (keras)
4. Understanding CNNs, Transfer Learning : Convolutional Neural Networks(CNNs) Transfer learning
5. Object Detection Algorithms
6. Image Segmentation & Attention Models
7. Deep Learning Tools : PyTorch & TensorFlow
8. Generative Adversial Networks (GANs)

References

- https://arxiv.org/pdf/1311.2524.pdf
- https://arxiv.org/pdf/1504.08083.pdf
- https://arxiv.org/pdf/1506.01497.pdf
- https://arxiv.org/pdf/1506.02640v5.pdf
- Deep Learning for Computer Vision- Standford
- https://pypi.org/project/face-recognition/

  - websites – machinelearningmastery.com, levelup.gitconnected.com, simplilearn.com, analyticsvidhya.com, towardsdatascience.com

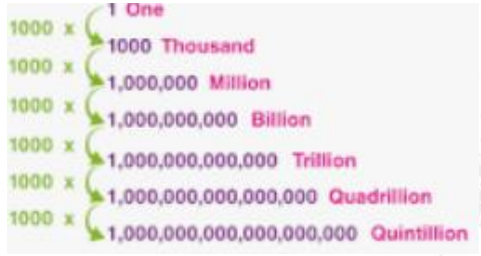
It is anticipated that the market for computer vision will approach **$41.11** billion by the year 2030, with a compound annual growth rate (CAGR) of **16.0%** between the years 2020 and 2030.

Allied Market Research

Computer Vision

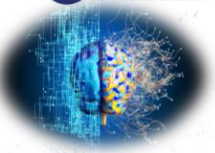
- Computer vision is one of the fields of artificial intelligence that trains and enables computers to understand the visual world.
- Computers can use digital images and deep learning models to accurately identify and classify objects and react to them.
- Computer vision in AI is dedicated to the development of automated systems that can interpret visual data (such as photographs or motion pictures) in the same manner as people do.
- The amount of data that we generate today is tremendous

  2.5 quintillion bytes ($2.5 * 10^{18}$) of data every single day.
- This growth in data has proven to be one of the driving factors behind the growth of computer vision



Billion of billion

History



1959

neurophysiologists started showing a cat a variety of sights in an effort to correlate a reaction in the animal's brain

1960's

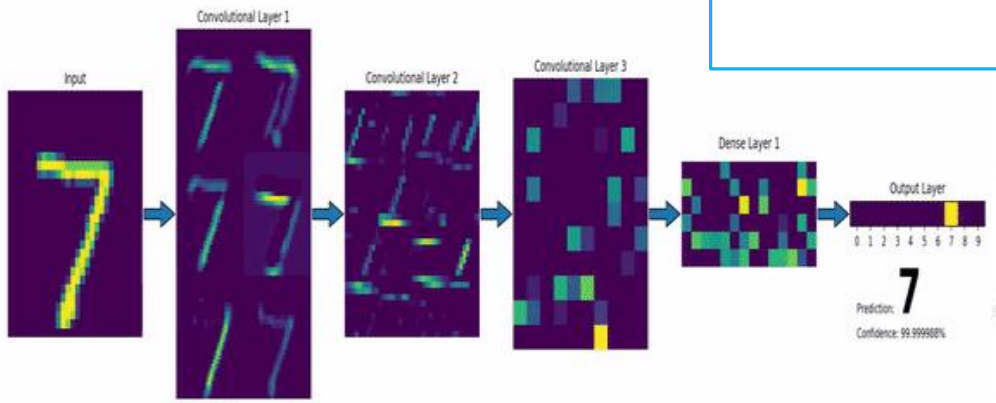
artificial intelligence (AI) emerged as an area of research, and the effort to address AI's inability to mimic human vision began.

1982

Neuroscientists demonstrated, vision operates hierarchically and presented techniques enabling computers to recognize edges, vertices, arcs, and other fundamental structures

2000

researchers were concentrating their efforts on object identification, and the industry saw the first-ever **real-time face recognition** solutions.



OpenCV

- Open cv is the most popular library in computer vision.
- It is originally written in C and C++, now it is available in python also.
- It is originally developed by intel.
- The library is a cross-platform open-source library. It is free to use.
- Open CV library is a highly optimized library with its main focus on real-time applications.
- Open CV is also used for Creating Face Recognition Systems.

OpenCV

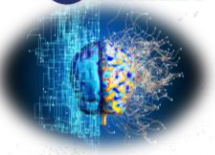
- The library has more than 2500 optimized algorithms.
  - which can be used to detect and recognize faces, identify objects, classify human actions using videos, tracking camera movements, tracking moving objects, extracting 3D models of objects, stitch images together to produce a high-resolution image of an entire scene,
  - find similar images from an image database, remove red eyes from images taken using flash, follow eye movements, etc
- It has around 47 thousand people of users community and an estimated number of downloads exceeding 18 million.
- Many big companies like google, amazon, Tesla, Microsoft, Honda, etc. uses Open cv to make their products better and more AI-driven.

OpenCV

- OpenCV uses machine learning algorithms to search for faces within a picture. Because faces are so complicated, there isn't one simple test that will tell you if it found a face or not. Instead, there are thousands of small patterns and features that must be matched. The algorithms break the task of identifying the face into thousands of smaller, bite-sized tasks, each of which is easy to solve. These tasks are also called classifiers.
- For something like a face, you might have 6,000 or more classifiers, all of which must match for a face to be detected (within error limits, of course). But therein lies the problem: for face detection, the algorithm starts at the top left of a picture and moves down across small blocks of data, looking at each block, constantly asking, "Is this a face? … Is this a face? … Is this a face?"
- Since there are 6,000 or more tests per block, you might have millions of calculations to do, which will grind your computer to a halt.

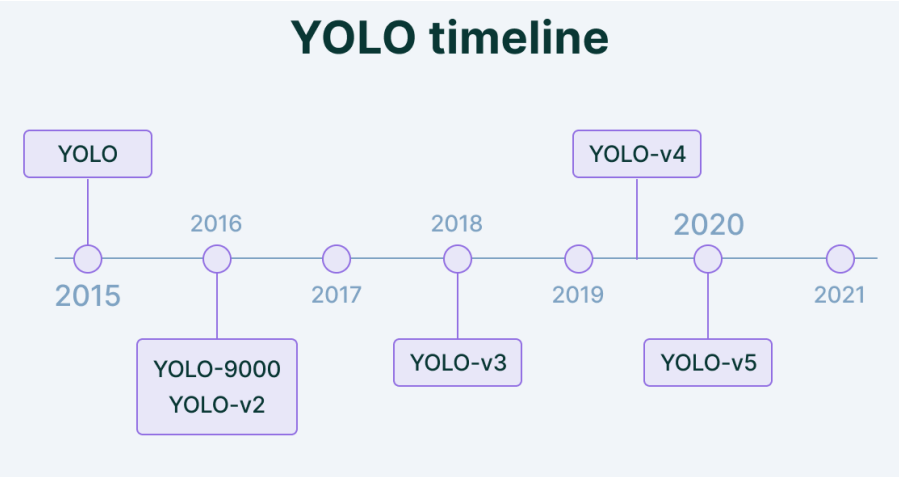
OpenCV

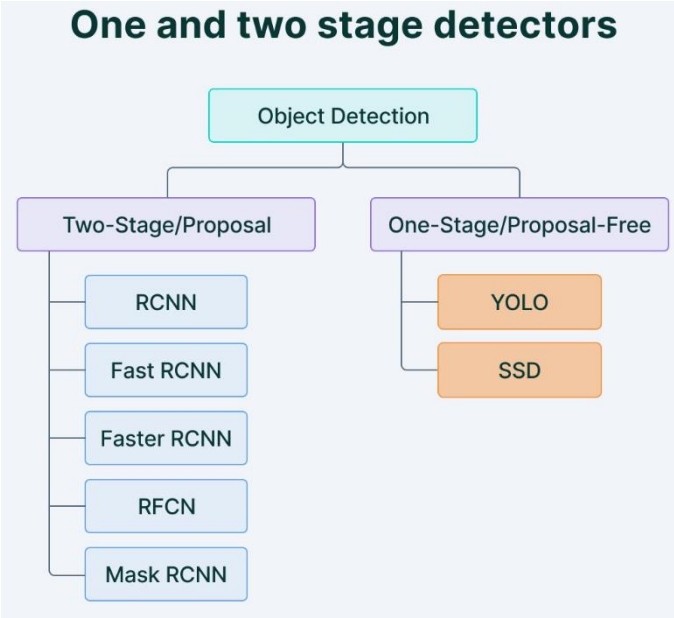
- To get around this, OpenCV uses cascades. What's a cascade? The best answer can be found in the dictionary: "a waterfall or series of waterfalls."
- Like a series of waterfalls, the OpenCV cascade breaks the problem of detecting faces into multiple stages. For each block, it does a very rough and quick test. If that passes, it does a slightly more detailed test, and so on. The algorithm may have 30 to 50 of these stages or cascades, and it will only detect a face if all stages pass.



OpenCV alternatives

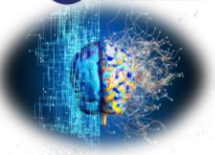
1) Microsoft Computer Vision API
2) AWS Rekognition
3) Google Cloud Vision API
4) Scikit-Image
5) SimpleCV
6) Azure Face API
7) DeepDream
8) IBM Watson Visual Recognition
9) Clarifi
10) DeepPy





Object Recognition

HUMANS CAN EASILY DETECT AND IDENTIFY OBJECT DETECTION PRESENT IN AN IMAGE. THE HUMAN VISUAL SYSTEM IS FAST AND ACCURATE AND CAN PERFORM COMPLEX TASKS LIKE IDENTIFYING MULTIPLE OBJECTS AND DETECTING OBSTACLES WITH LITTLE CONSCIOUS THOUGHT. WITH THE AVAILABILITY OF LARGE AMOUNTS OF DATA, FASTER GPUS, AND BETTER ALGORITHMS, WE CAN NOW EASILY TRAIN COMPUTERS TO DETECT AND CLASSIFY MULTIPLE OBJECTS WITHIN AN IMAGE WITH HIGH ACCURACY.

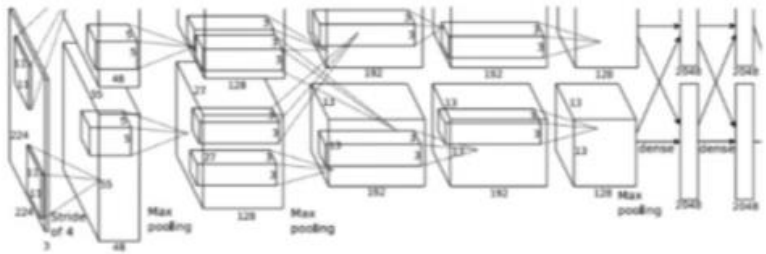


Computer Vision

- Computer vision is an interdisciplinary field that has been gaining huge amounts of traction in the recent years and self-driving cars have taken centre stage.
- Another integral part of computer vision is object detection.
- Object detection aids in pose estimation, vehicle detection, surveillance etc.
- The difference between object detection algorithms and classification algorithms is that in detection algorithms, we try to draw a bounding box around the object of interest to locate it within the image.
- Also, you might not necessarily draw just one bounding box in an object detection case, there could be many bounding boxes representing different objects of interest within the image and you would not know how many beforehand.

Computer Vision

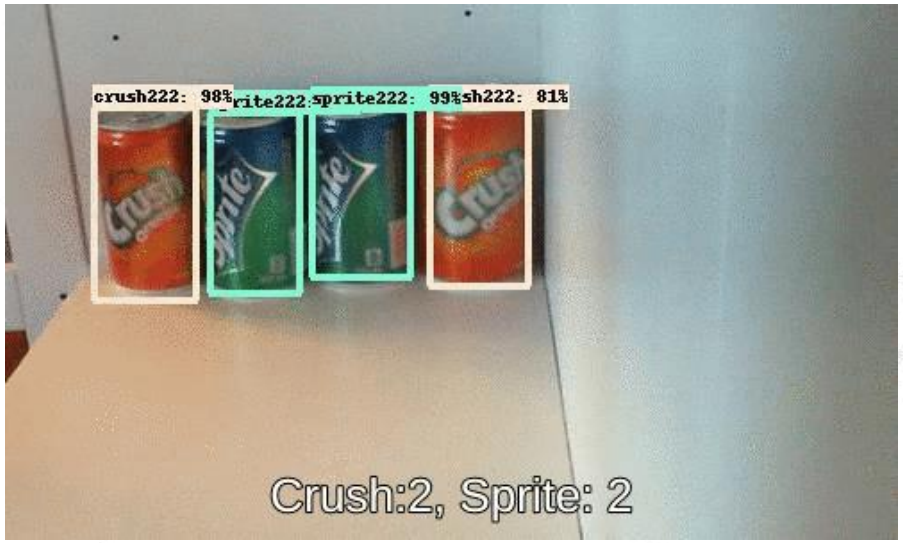
- The major reason why you cannot proceed with this problem by building a standard convolutional network followed by a fully connected layer is that, the length of the output layer is variable — not constant, this is because the number of occurrences of the objects of interest is not fixed.
- A naive approach to solve this problem would be to take different regions of interest from the image, and use a CNN to classify the presence of the object within that region. The problem with this approach is that the objects of interest might have different spatial locations within the image and different aspect ratios. Hence, you would have to select a huge number of regions and this could computationally blow up.
- Therefore, algorithms like R-CNN, YOLO etc have been developed to find these occurrences and find them fast.

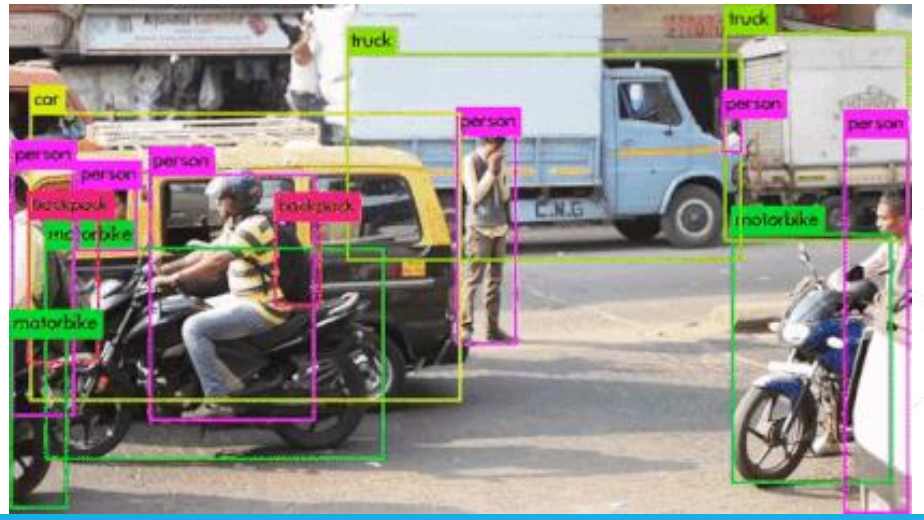


CAT: (x, y, w, h)

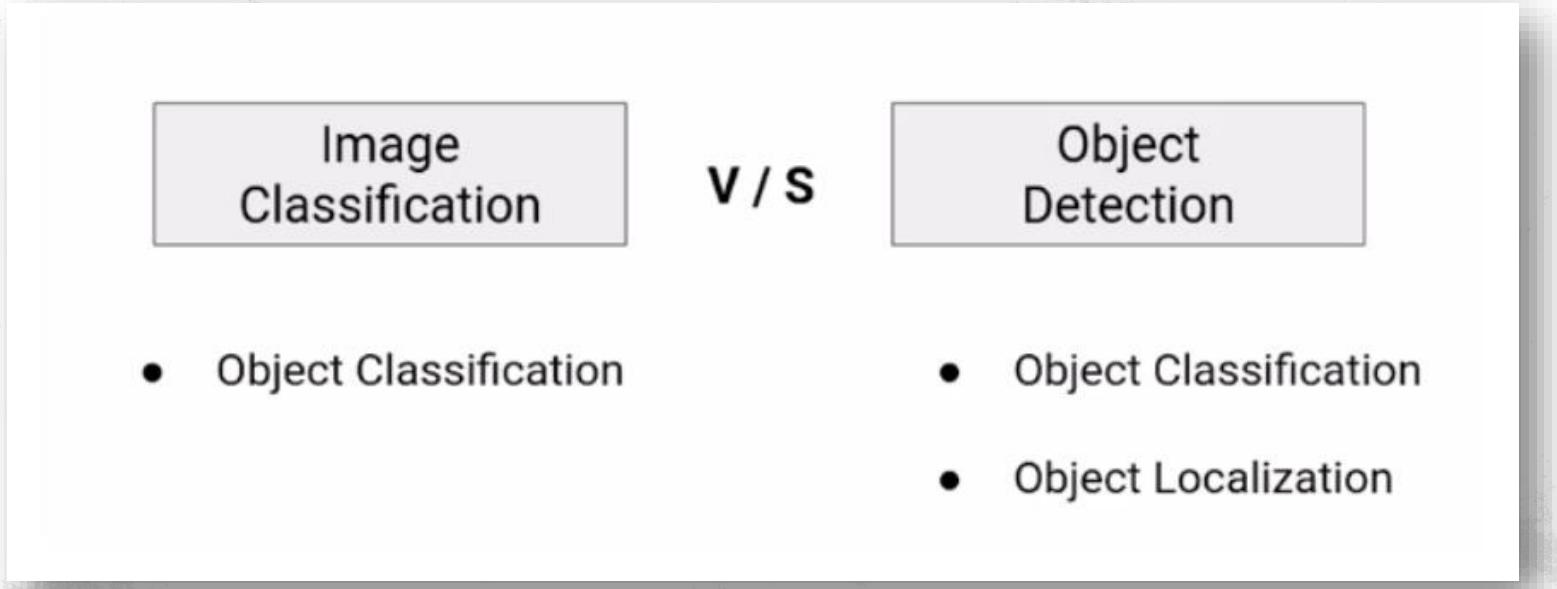
Object Recognition

- Object recognition is a general term to describe a collection of related computer vision tasks that involve identifying objects in digital photographs.





- Object Detection - we have to classify the objects in the image and also locate where these objects are present in the image



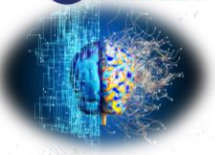


Image Classification

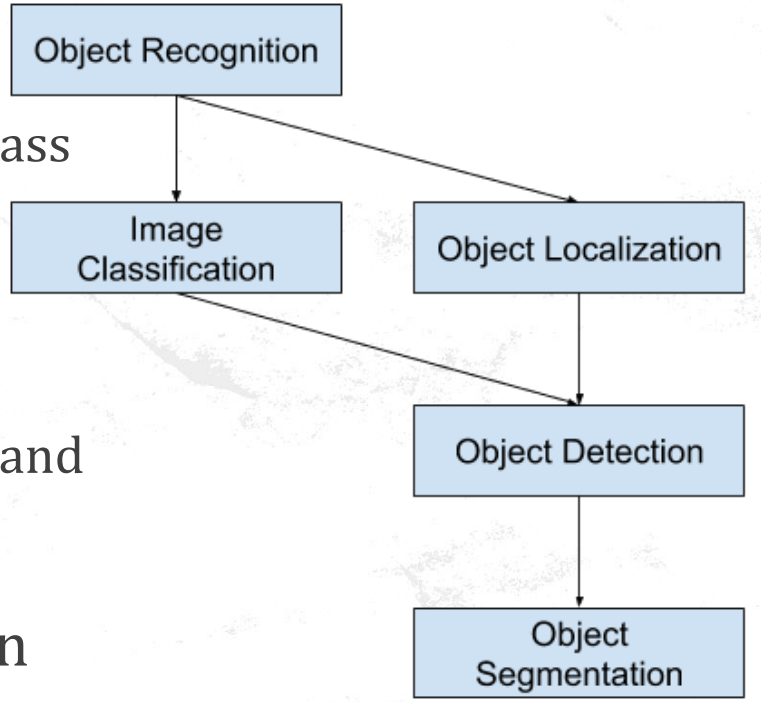
Image classification involves predicting the class of one object in an image.

- ***Object localization*** refers to identifying the location of one or more objects in an image and drawing abounding box around their extent.
- ***Object detection*** combines these two tasks and localizes and classifies one or more objects in an image.

> *"... we will be using the term object recognition broadly to encompass both image classification (a task requiring an algorithm to determine what object classes are present in the image) as well as object detection (a task requiring an algorithm to localize all objects present in the image)*
>
> — ImageNet Large Scale Visual Recognition Challenge, 2015.

Image Classification

- **Image Classification**: Predict the type or class of an object in an image.
  - *Input*: An image with a single object, such as a photograph.
  - *Output*: A class label (e.g. one or more integers that are mapped to class labels).
- **Object Localization**: Locate the presence of objects in an image and indicate their location with a bounding box.
  - *Input*: An image with one or more objects, such as a photograph.
  - *Output*: One or more bounding boxes (e.g. defined by a point, width, and height).
- **Object Detection**: Locate the presence of objects with a bounding box and types or classes of the located objects in an image.
  - *Input*: An image with one or more objects, such as a photograph.
  - *Output*: One or more bounding boxes (e.g. defined by a point, width, and height), and a class label for each bounding box.



ImageNet Large Scale Visual Recognition Challenge

- Most of the recent innovations in image recognition problems have come as part of participation in the ImageNet Large Scale Visual Recognition Challenge(ILSVRC) tasks.
- This is an annual academic competition with a separate challenge for each of these three problem types, with the intent of fostering independent and separate improvements at each level that can be leveraged more broadly.

https://arxiv.org/abs/1409.0575

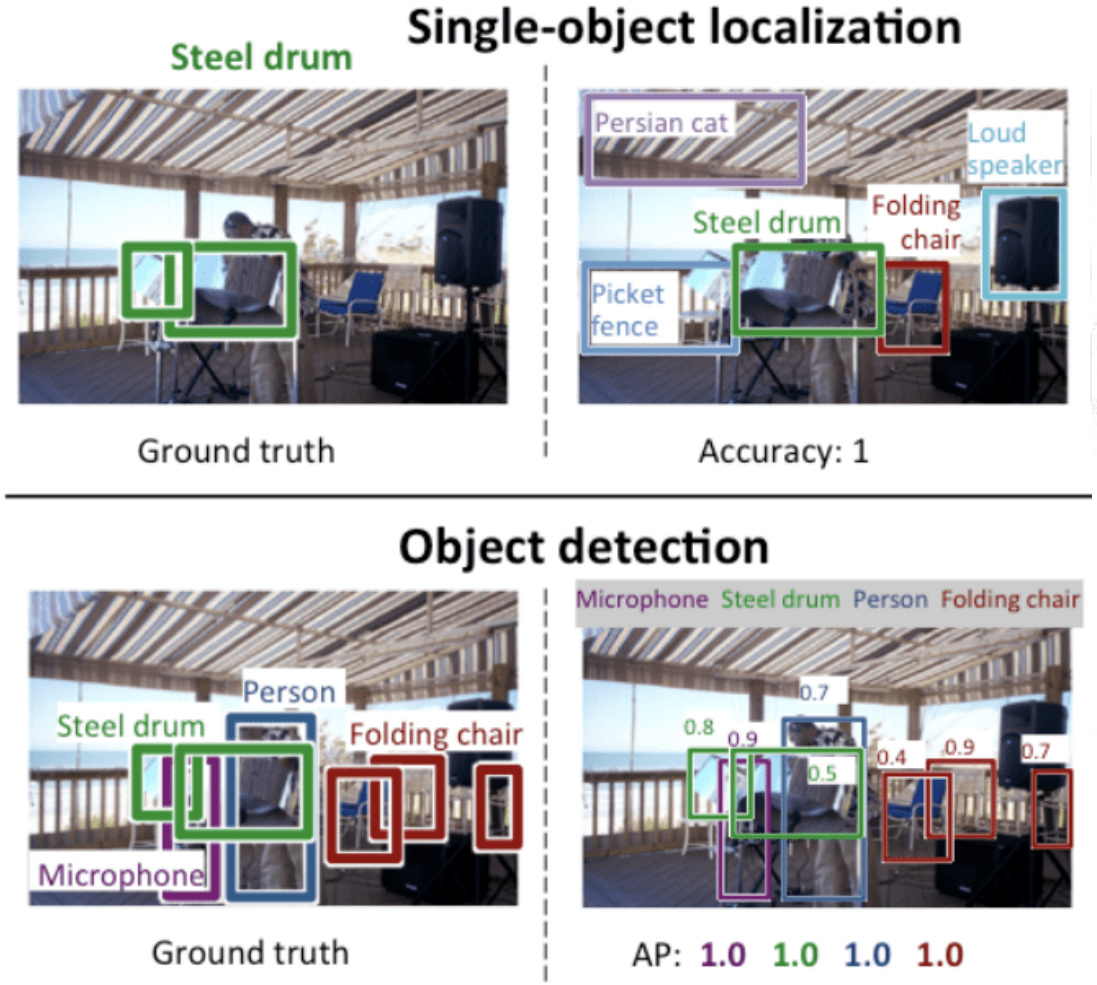
ImageNet Large Scale Visual Recognition Challenge

List of the three corresponding task types below taken from the 2015 ILSVRC review paper:

- **Image classification**: Algorithms produce a list of object categories present in the image.
- **Single-object localization**: Algorithms produce a list of object categories present in the image, along with an axis-aligned bounding box indicating the position and scale of one instance of each object category.
- **Object detection**: Algorithms produce a list of object categories present in the image along with an axis-aligned bounding box indicating the position and scale of every instance of each object category.

ILSVRC

- The performance of a model for image classification is evaluated using the mean classification error across the predicted class labels.
- The performance of a model for single-object localization is evaluated using the distance between the expected and predicted bounding box for the expected class.
- Whereas the performance of a model for object recognition is evaluated using the precision and recall across each of the best matching bounding boxes for the known objects in the image.



R-CNN Model Family

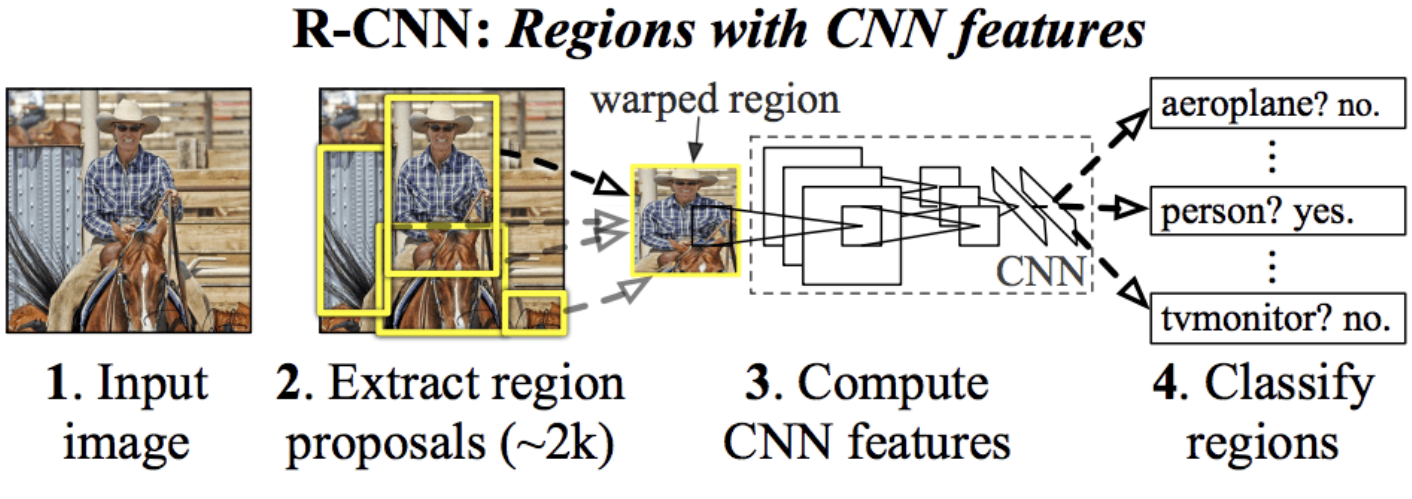
- The R-CNN family of methods refers to the R-CNN, which may stand for "*Regions with CNN Features*" or "*Region-Based Convolutional Neural Network*," developed by Ross Girshick, et al.
- This includes the techniques R-CNN, Fast R-CNN, and Faster-RCNN designed and demonstrated for object localization and object recognition.
- The R-CNN was described in the 2014 paper by Ross Girshick, et al. from UC Berkeley titled "Rich feature hierarchies for accurate object detection and semantic segmentation."
- It may have been one of the first large and successful application of convolutional neural networks to the problem of object localization, detection, and segmentation.

R-CNN

Their proposed R-CNN model is comprised of three modules; they are:

- **Module 1: Region Proposal**. Generate and extract category independent region proposals, e.g. candidate bounding boxes.
- **Module 2: Feature Extractor**. Extract feature from each candidate region, e.g. using a deep convolutional neural network.
- **Module 3: Classifier**. Classify features as one of the known class, e.g. linear SVM classifier model.

The architecture of the model is summarized in the image below, taken from the paper.



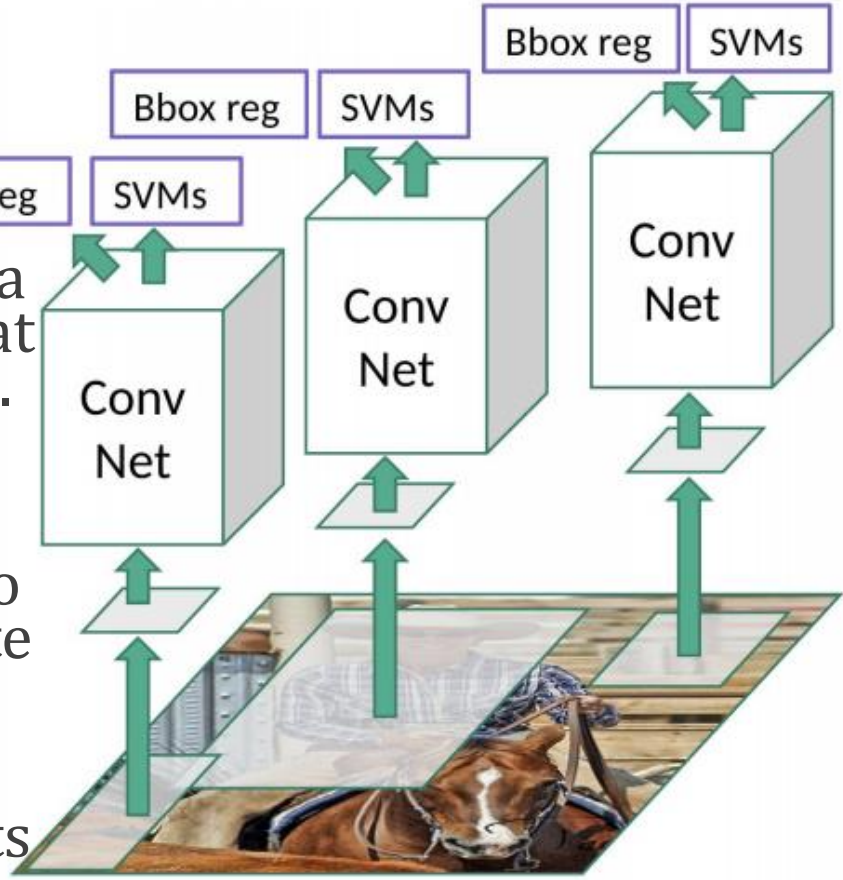
R-CNN

- To bypass the problem of selecting a huge number of regions, Ross Girshick et al. proposed a method where we use selective search to extract just 2000 regions from the image and he called them region proposals.
- Therefore, now, instead of trying to classify a huge number of regions, you can just work with 2000 regions. These 2000 region proposals are generated using the selective search algorithm which is written below.

Selective Search (link) :
1. Generate initial sub-segmentation, we generate many candidate regions
2. Use greedy algorithm to recursively combine similar regions into larger ones
3. Use the generated regions to produce the final candidate region proposals

R-CNN

- The 2000 candidate region proposals are warped into a square and fed into a convolutional neural network that produces a 4096-dimensional feature vector as output.
- The CNN acts as a feature extractor and the output dense layer consists of the features extracted from the image and the extracted features are fed into an SVM to classify the presence of the object within that candidate region proposal.
- In addition to predicting the presence of an object within the region proposals, the algorithm also predicts four values which are offset values to increase the precision of the bounding box.
- For example, given a region proposal, the algorithm would have predicted the presence of a person but the face of that person within that region proposal could've been cut in half. Therefore, the offset values help in adjusting the bounding box of the region proposal

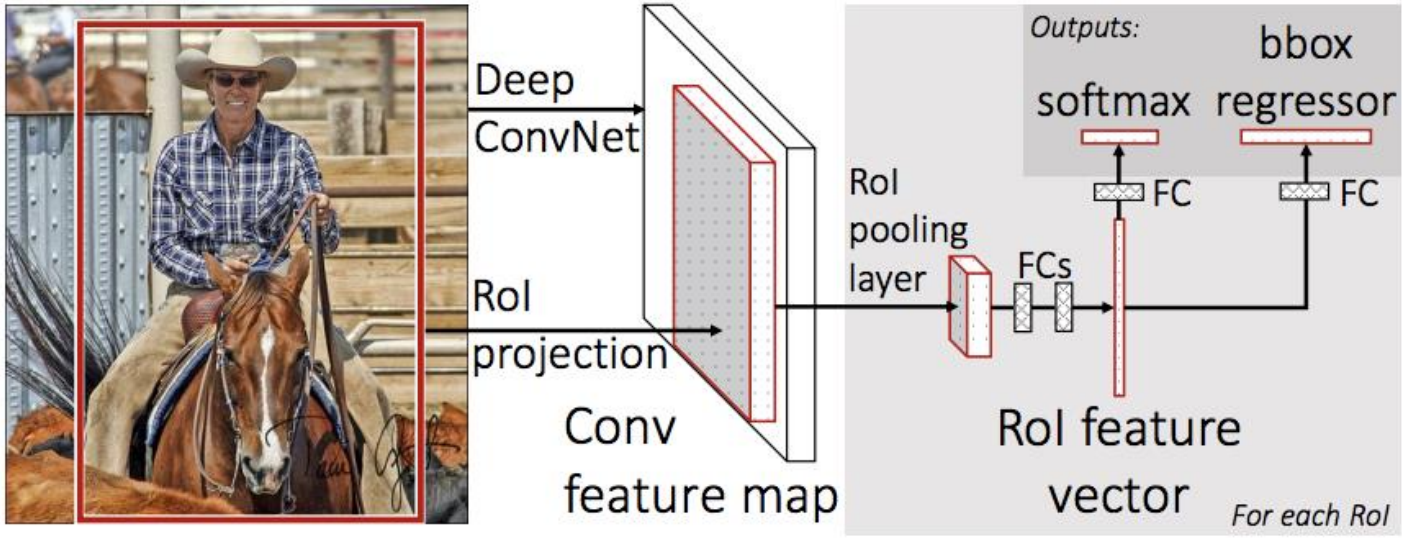


Problems with R-CNN

- It still takes a huge amount of time to train the network as you would have to classify 2000 region proposals per image.
- It cannot be implemented real time as it takes around 47 seconds for each test image.
- The selective search algorithm is a fixed algorithm. Therefore, no learning is happening at that stage. This could lead to the generation of bad candidate region proposals.

Fast R-CNN

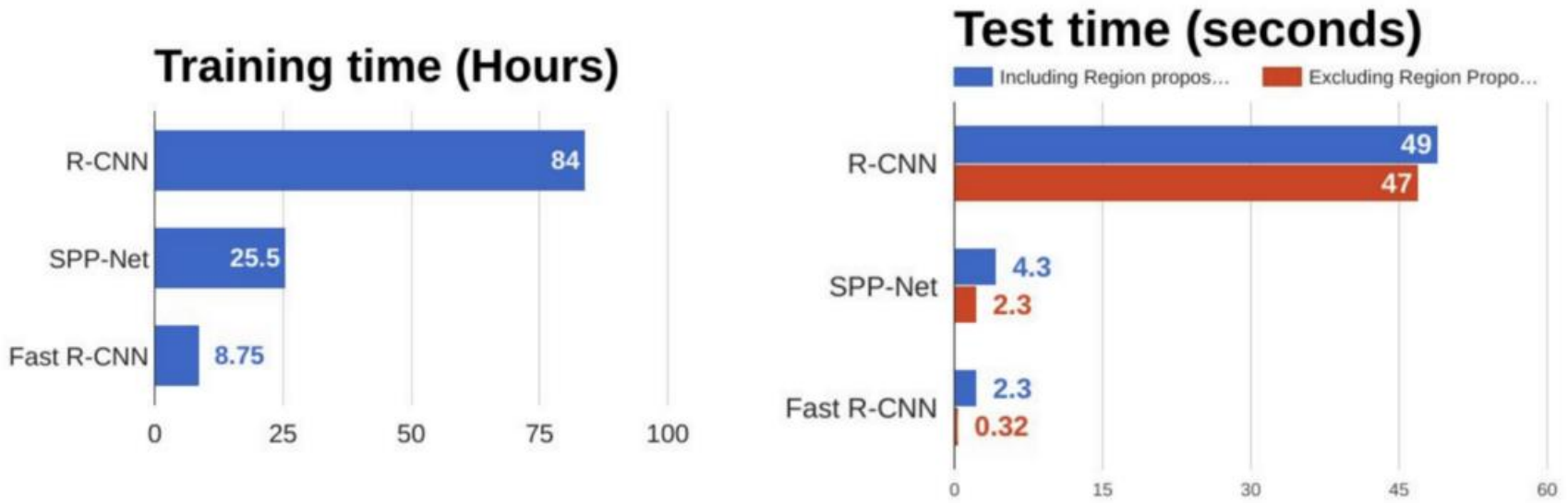
The same author of the previous paper(R-CNN) solved some of the drawbacks of R-CNN to build a faster object detection algorithm and it was called Fast R-CNN.



- The approach is similar to the R-CNN algorithm. But, instead of feeding the region proposals to the CNN, we feed the input image to the CNN to generate a convolutional feature map.
- From the convolutional feature map, we identify the region of proposals and warp them into squares and by using a RoI pooling layer we reshape them into a fixed size so that it can be fed into a fully connected layer.
- From the RoI feature vector, we use a softmax layer to predict the class of the proposed region and also the offset values for the bounding box.

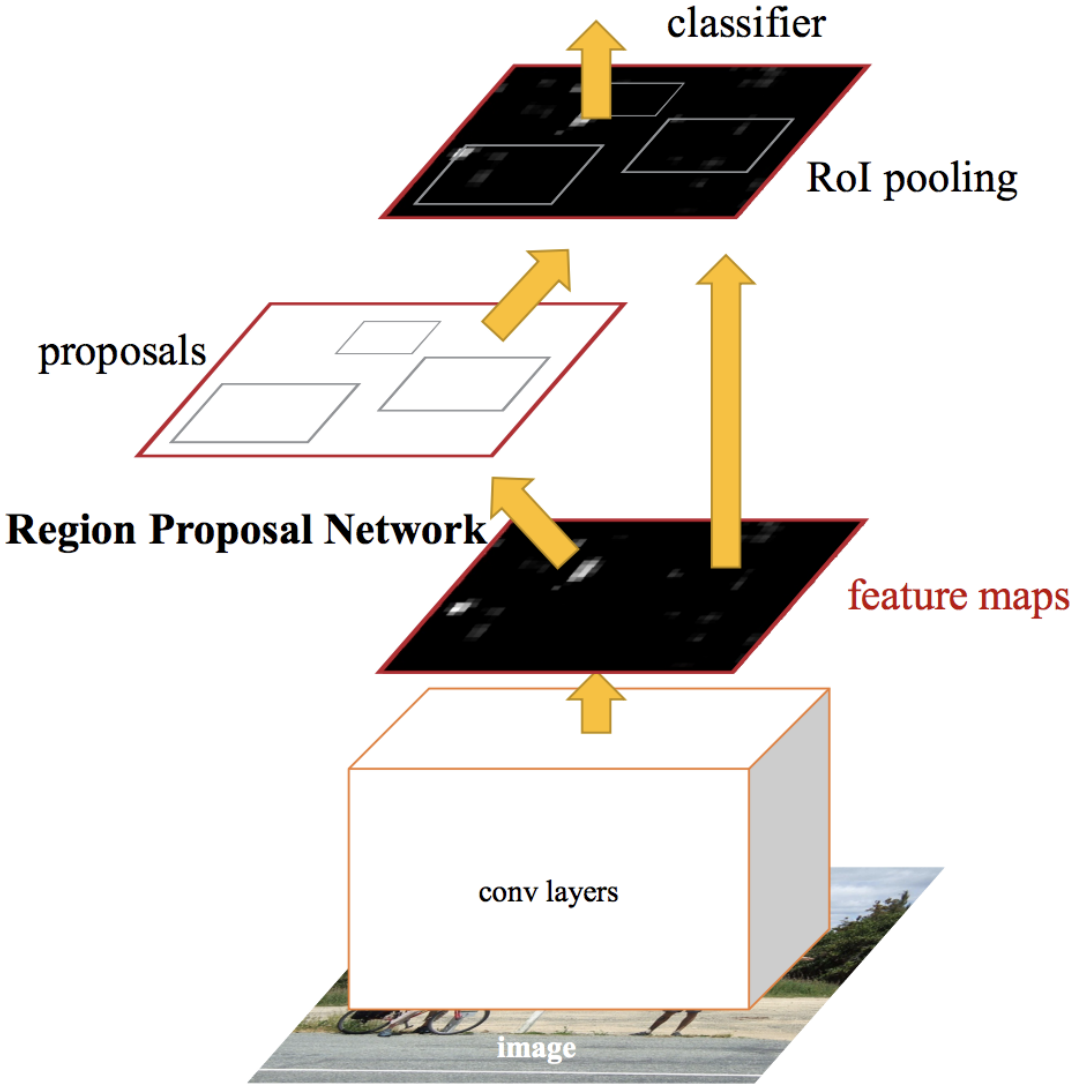
Fast R-CNN

- The reason “Fast R-CNN” is faster than R-CNN is because you don’t have to feed 2000 region proposals to the convolutional neural network every time. Instead, the convolution operation is done only once per image and a feature map is generated from it.
-



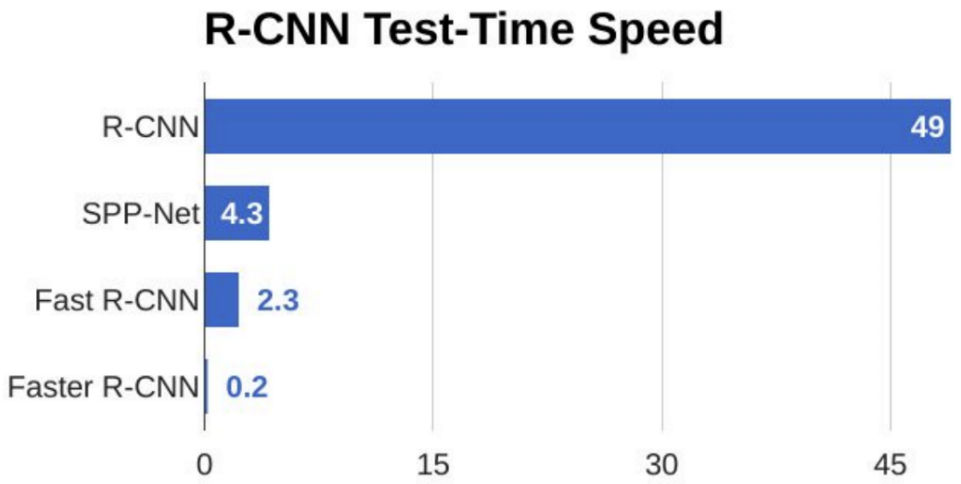
Faster R-CNN

- Both of the above algorithms(R-CNN & Fast R-CNN) uses selective search to find out the region proposals. Selective search is a slow and time-consuming process affecting the performance of the network.
- Therefore, Shaoqing Ren et al. came up with an object detection algorithm that eliminates the selective search algorithm and lets the network learn the region proposals.

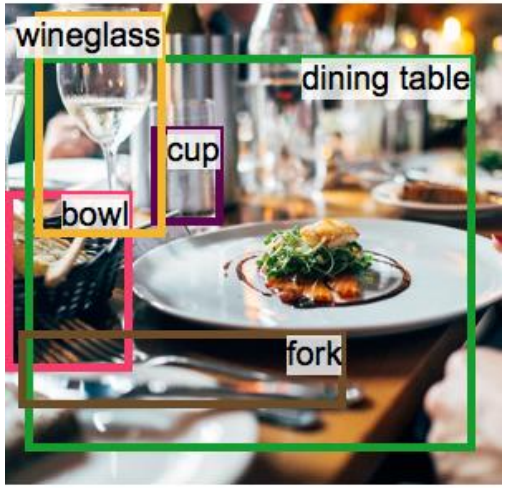


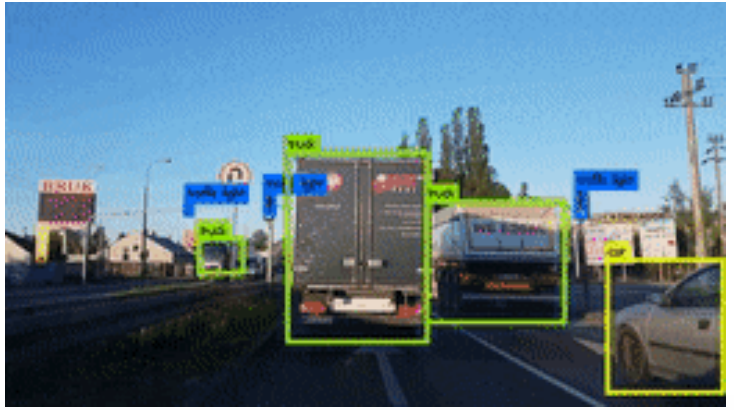
Faster R-CNN

- Similar to Fast R-CNN, the image is provided as an input to a convolutional network which provides a convolutional feature map.
- Instead of using selective search algorithm on the feature map to identify the region proposals, a separate network is used to predict the region proposals.
- The predicted region proposals are then reshaped using a RoI pooling layer which is then used to classify the image within the proposed region and predict the offset values for the bounding boxes.



WOLFRAM MATHEMATICA

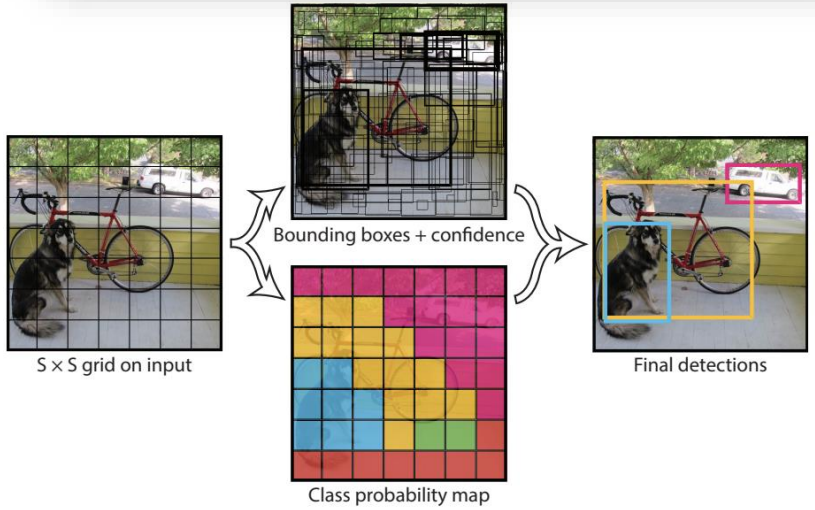




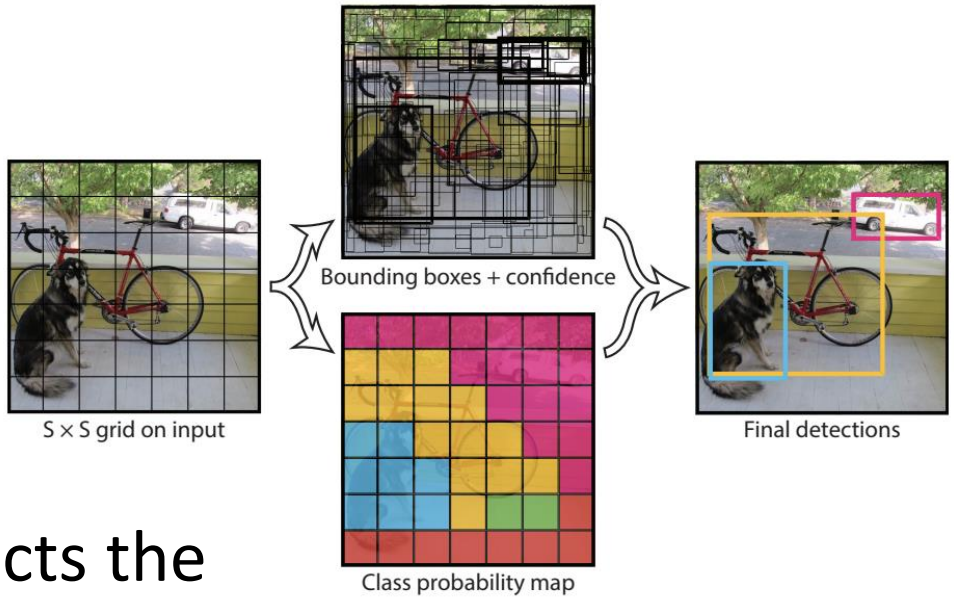
YOLO

You Only Look Once

YOLO IS AN OBJECT DETECTION ALGORITHM MUCH DIFFERENT FROM THE REGION BASED ALGORITHMS



YOLO



- YOLO (You Only Look Once) is an object detection algorithm much different from the region based algorithms
- In YOLO a single convolutional network predicts the bounding boxes and the class probabilities for these boxes.
- YOLO take an image and split it into an SxS grid, within each of the grid we take m bounding boxes.
- For each of the bounding box, the network outputs a class probability and offset values for the bounding box.
- The bounding boxes having the class probability above a threshold value is selected and used to locate the object within the image.

YOLO

- YOLO is orders of magnitude faster(45 frames per second) than other object detection algorithms.
- The limitation of YOLO algorithm is that it struggles with small objects within the image, for example it might have difficulties in detecting a flock of birds.
- This is due to the spatial constraints of the algorithm.

- https://huggingface.co/spaces/sarwansingh/yolov8

Source : fcakyon/yolov8-segmentation

Input Image

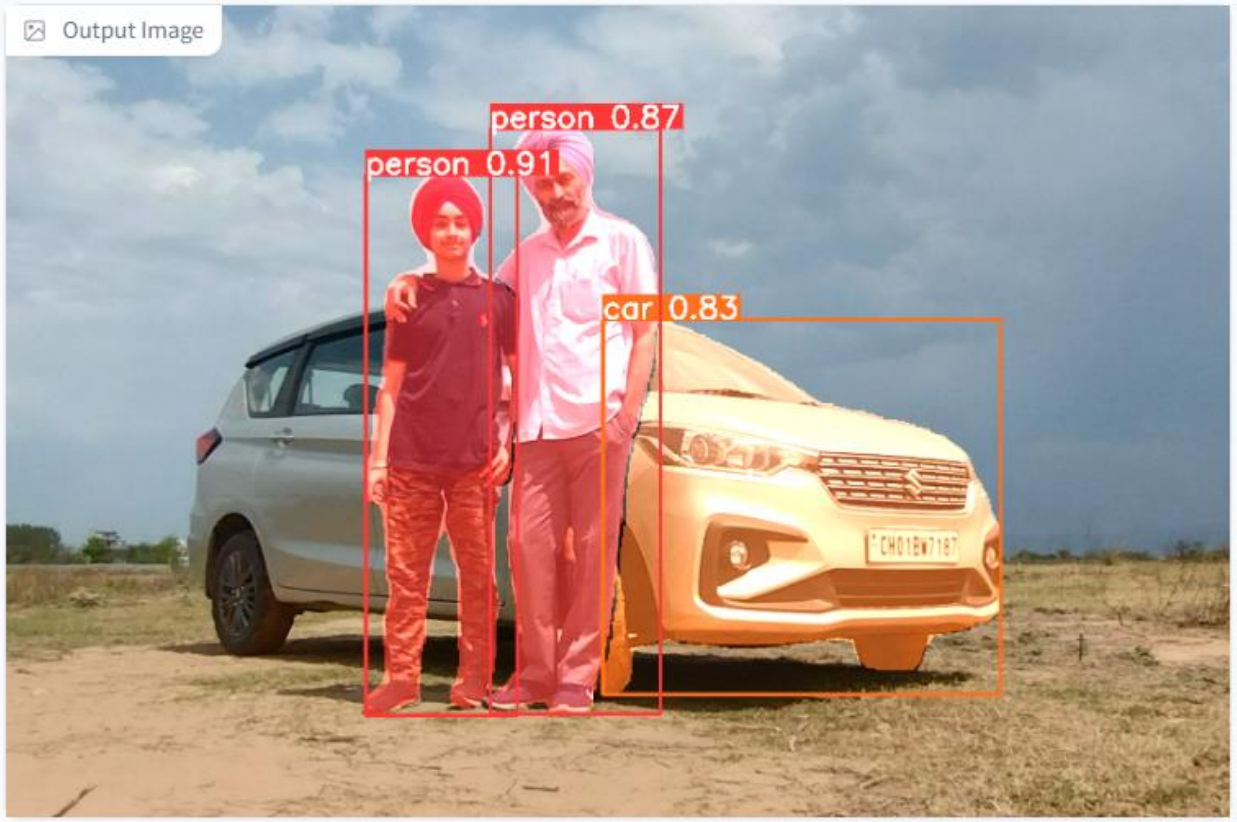
Model type

yolov8m-seg.pt

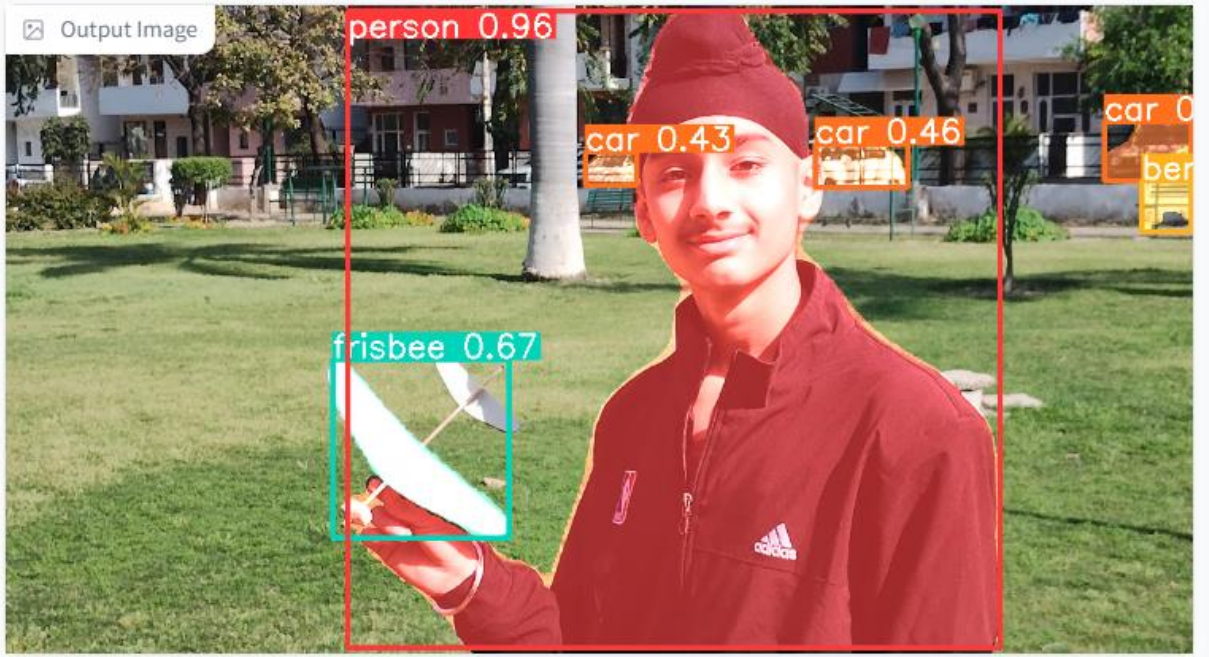
Image Size

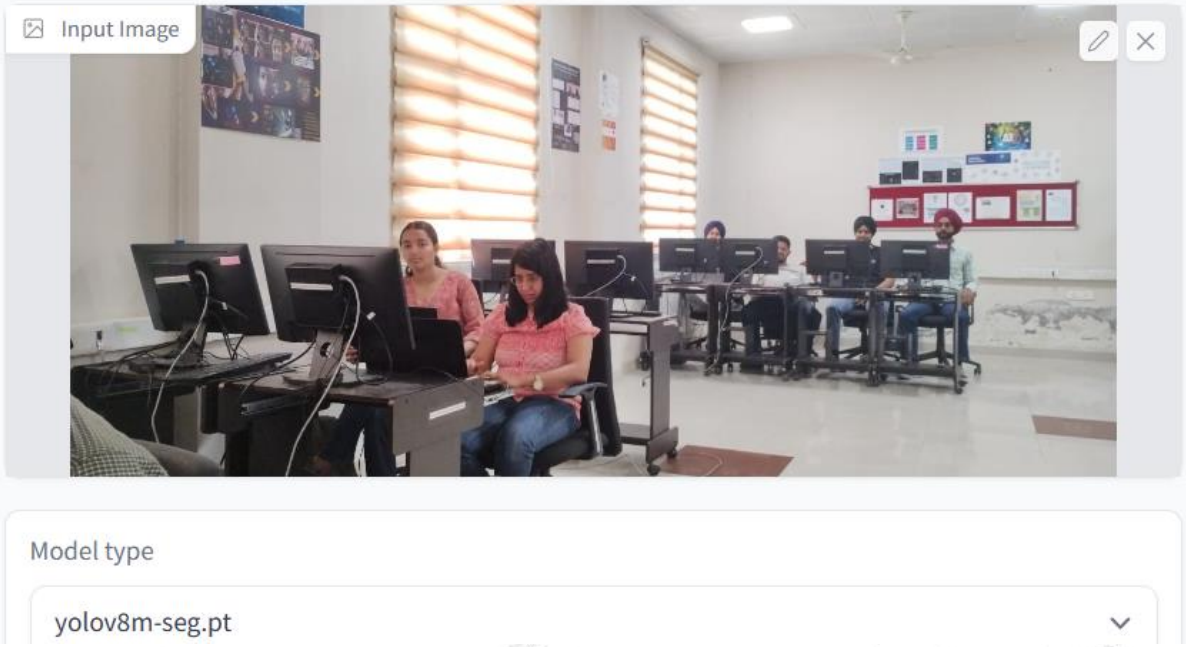
640

Output Image

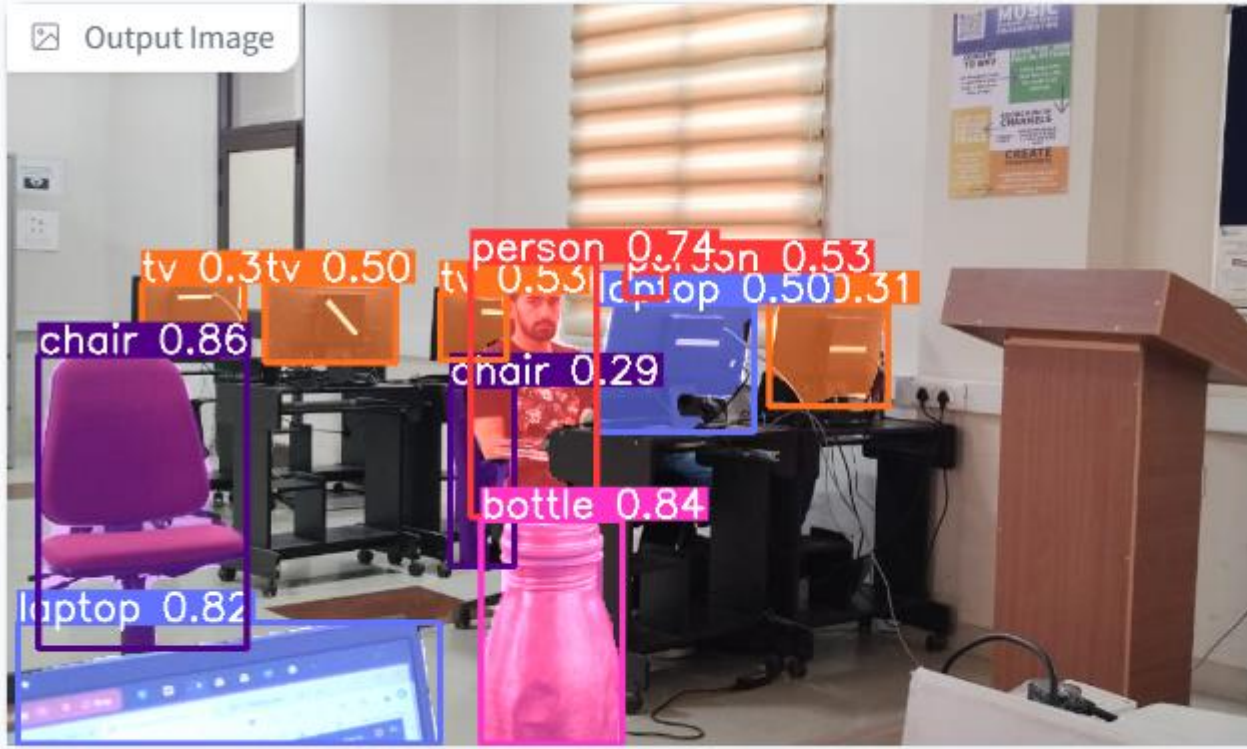



Input Image
Model type
yolov8m-seg.pt


Output Image
person 0.96
car 0.43
car 0.46
frisbee 0.67


Input Image
Model type
yolov8m-seg.pt


Output Image
tv 0.75
laptop 0.62
person 0.84
person 0.91
person 0.77
person 0.34


Output Image
tv 0.3tv 0.50
person 0.74
chair 0.86
chair 0.29
bottle 0.84
laptop 0.82


Input Image
Model type


Output Image
refrigerator 0.55
tv 0.4
tv 0.55
person
person 0.70
chair 0.58
chair 0.91
0.42
0.26
chair 0.35

Ultralytics YOLOv8 Segmentation Demo