Khwopa
College of
Engineering

OBJECT RECOGNITION

PRESENTATION BY: BIBEK SHYAMA (074/BCT/015)

What is object recognition ?

- Object recognition is the area of AI concerned with the abilities to recognize an object in a images or in videos.
- Object recognition is a key output of deep learning and machine learning algorithms.

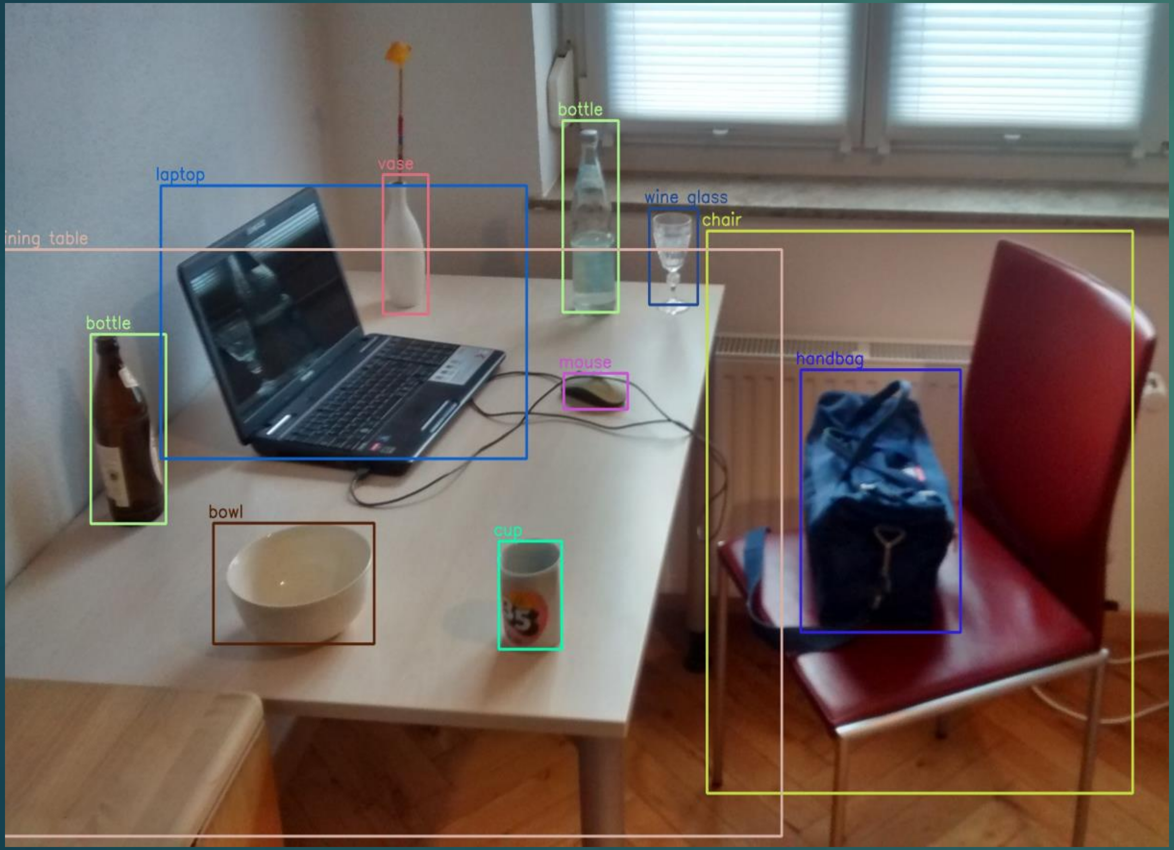


Fig 1 : Object recognition in image

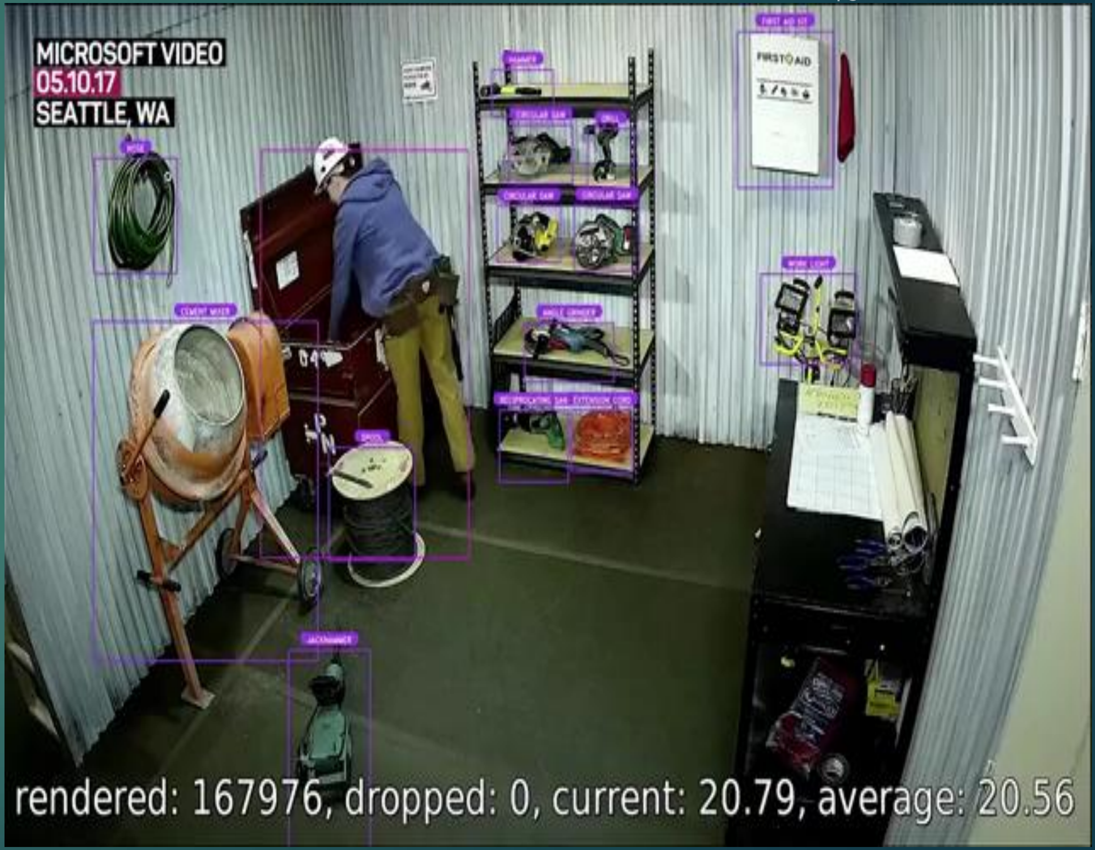

MICROSOFT VIDEO
05.10.17
SEATTLE, WA
rendered: 167976, dropped: 0, current: 20.79, average: 20.56

Fig 2 : Object recognition in video

How object recognition works ?

- 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.
- Object Segmentation : Instances of recognized objects are indicated by highlighting the specific pixels of the object instead of coarse bounding box.

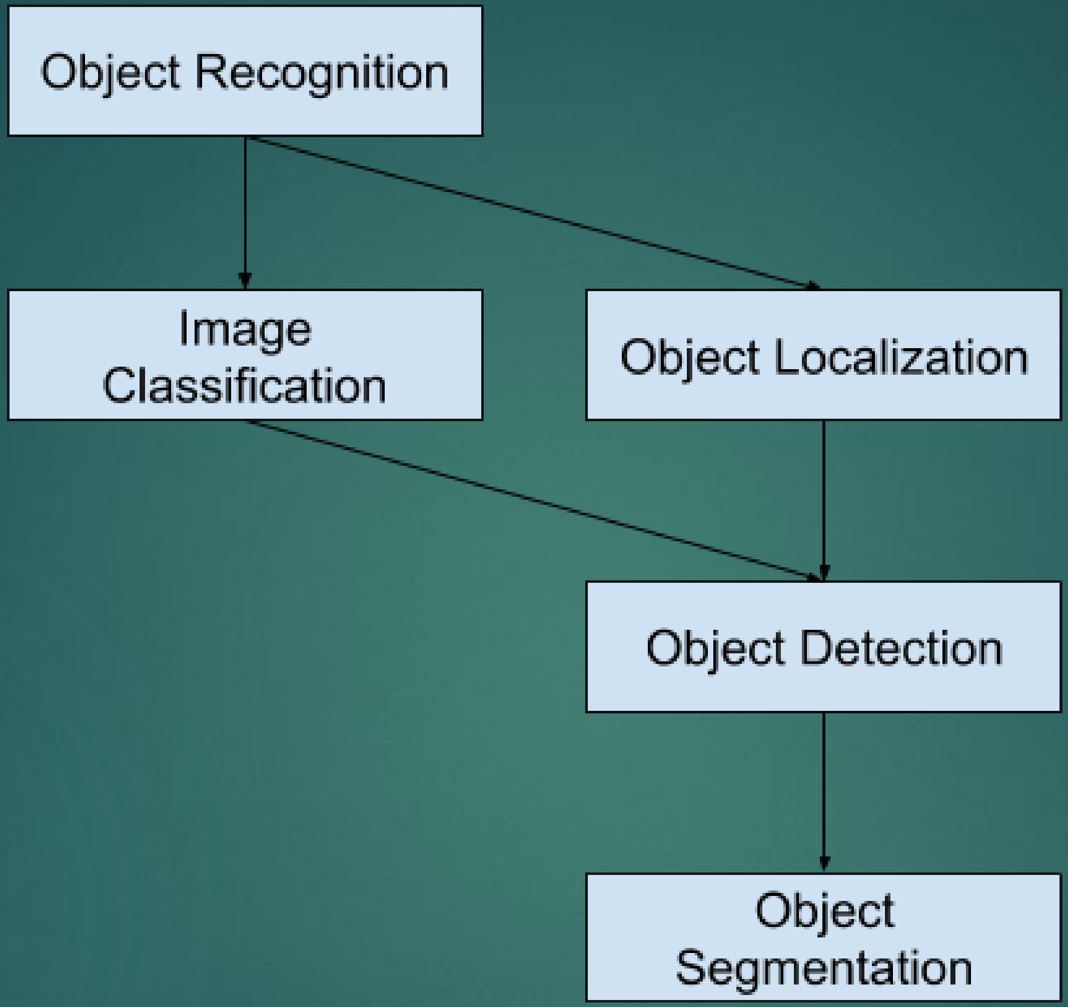


Fig 3: Overview of Object Recognition

How to achieve Object Recognition ?

- We can use a variety of approaches for object recognition. Recently, techniques in Machine Learning (ML) and Deep Learning have become popular approaches to object recognition problems. Both techniques learn to identify objects in images, but they differ in their execution.

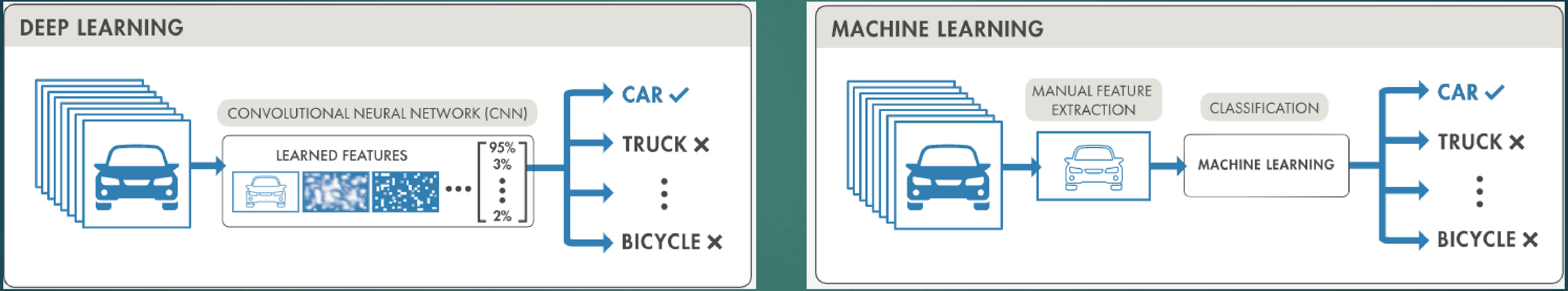


Fig 4: Deep learning and Machine learning techniques for Object Recognition

Object Recognition Using Machine Learning

- Techniques for Object Recognition using Machine Learning :
  - HOG feature extraction with an SVM machine learning model
  - Bags-of-Words models with features such as SURF and MSER
  - The Viola-Jon algorithm, which can be used to recognize a variety of objects, including faces and upper bodies.
- Machine Learning Workflow
  - To perform object recognition using a standard machine learning approach, we start with a collection of images (or video), and select the relevant features in each image. For example, a feature extraction algorithm might extract edge or corner features that can be used to differentiate between classes in your data.
  - These features are added to a machine learning model, which will separate these features into their distinct categories, and then use this information when analyzing and classifying new objects.
  - We can use a variety of machine learning algorithms and feature extraction methods, which offer many combinations to create an accurate object recognition model.

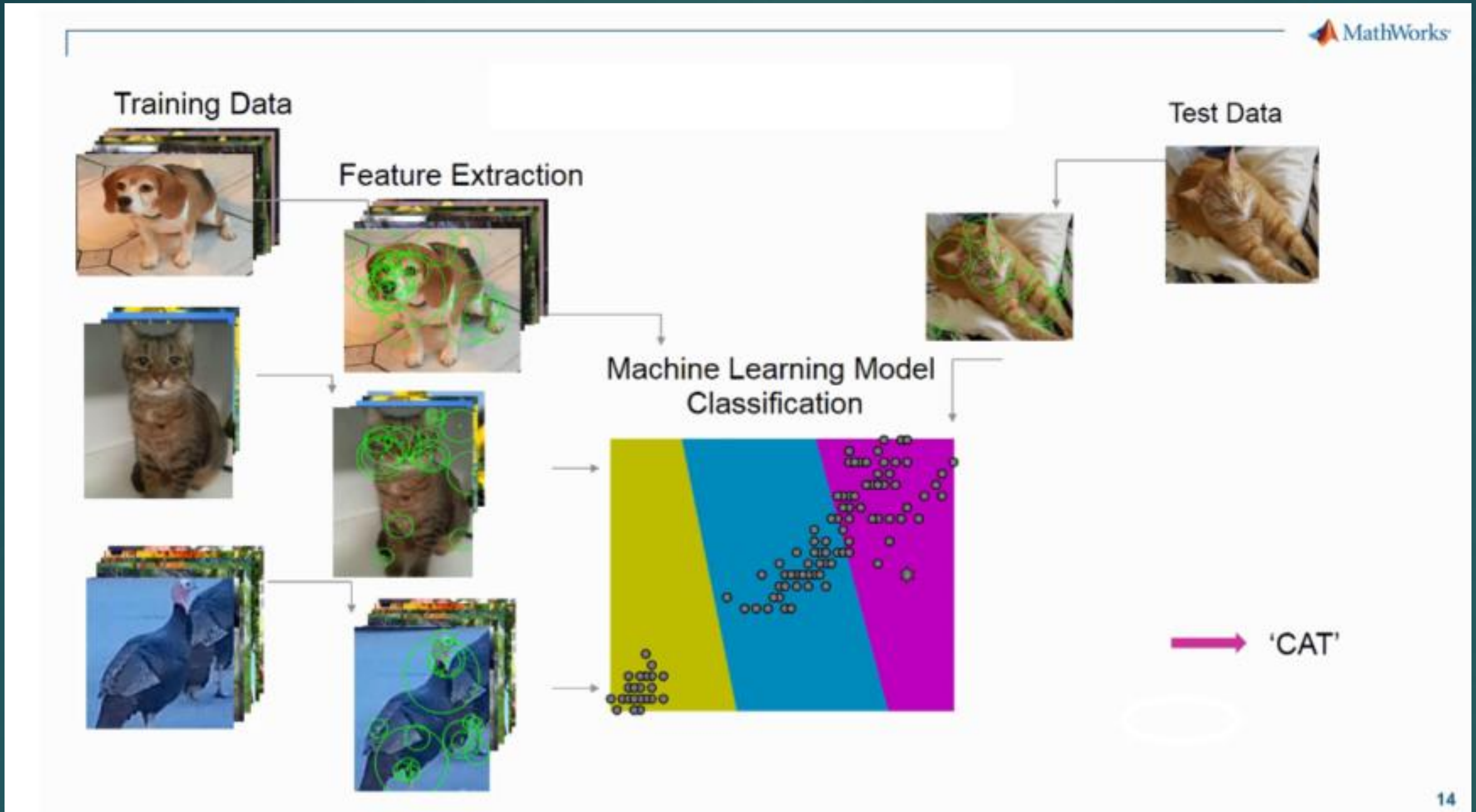


Fig : Machine learning workflow

Object Recognition Using Deep Learning

- Deep learning models such as convolutional neural networks (CNN) 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.

There are two approaches to performing object recognition using deep learning:

- **Training a model from scratch**: To train a deep network from scratch, we 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, and you need to set up the layers and weights in the CNN.
- **Using a pretrained deep learning model**: Most deep learning applications use the transfer learning approach, a process that involves fine-tuning a pretrained model. You start with an existing network, such as AlexNet or GoogLeNet, and feed in new data containing previously unknown classes. 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.

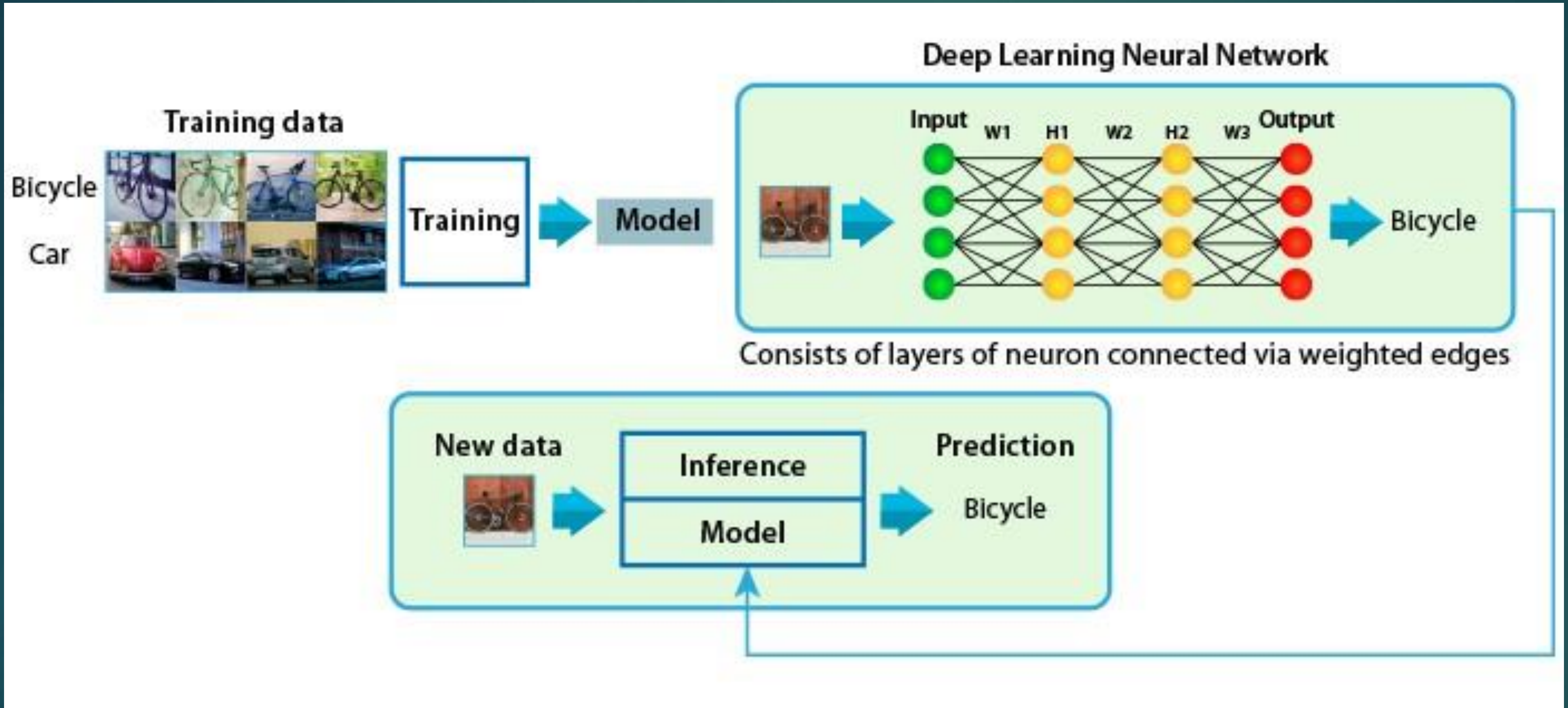


Fig 5: Deep Learning Workflow

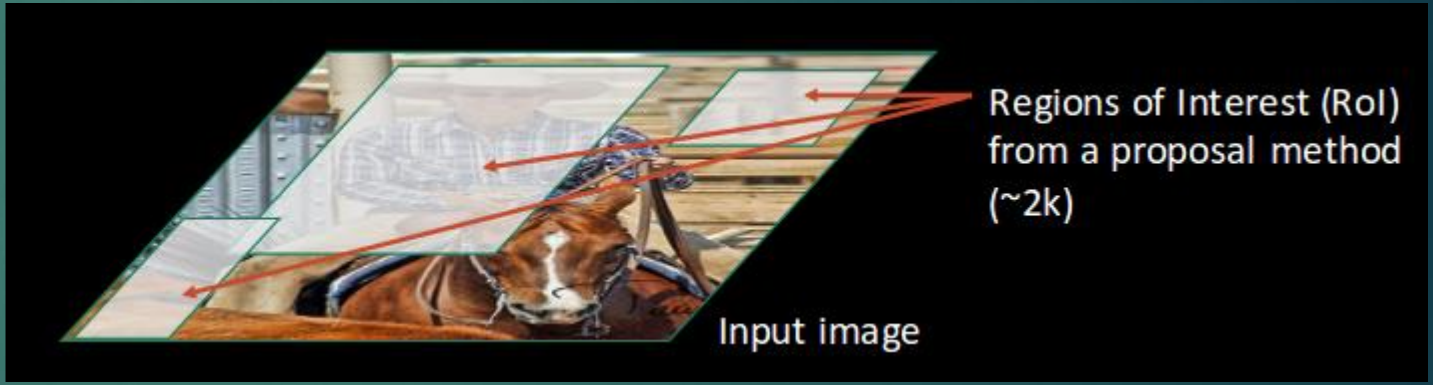
Models

- R-CNN model
  - R-CNN
  - Fast R-CNN
  - Faster R-CNN

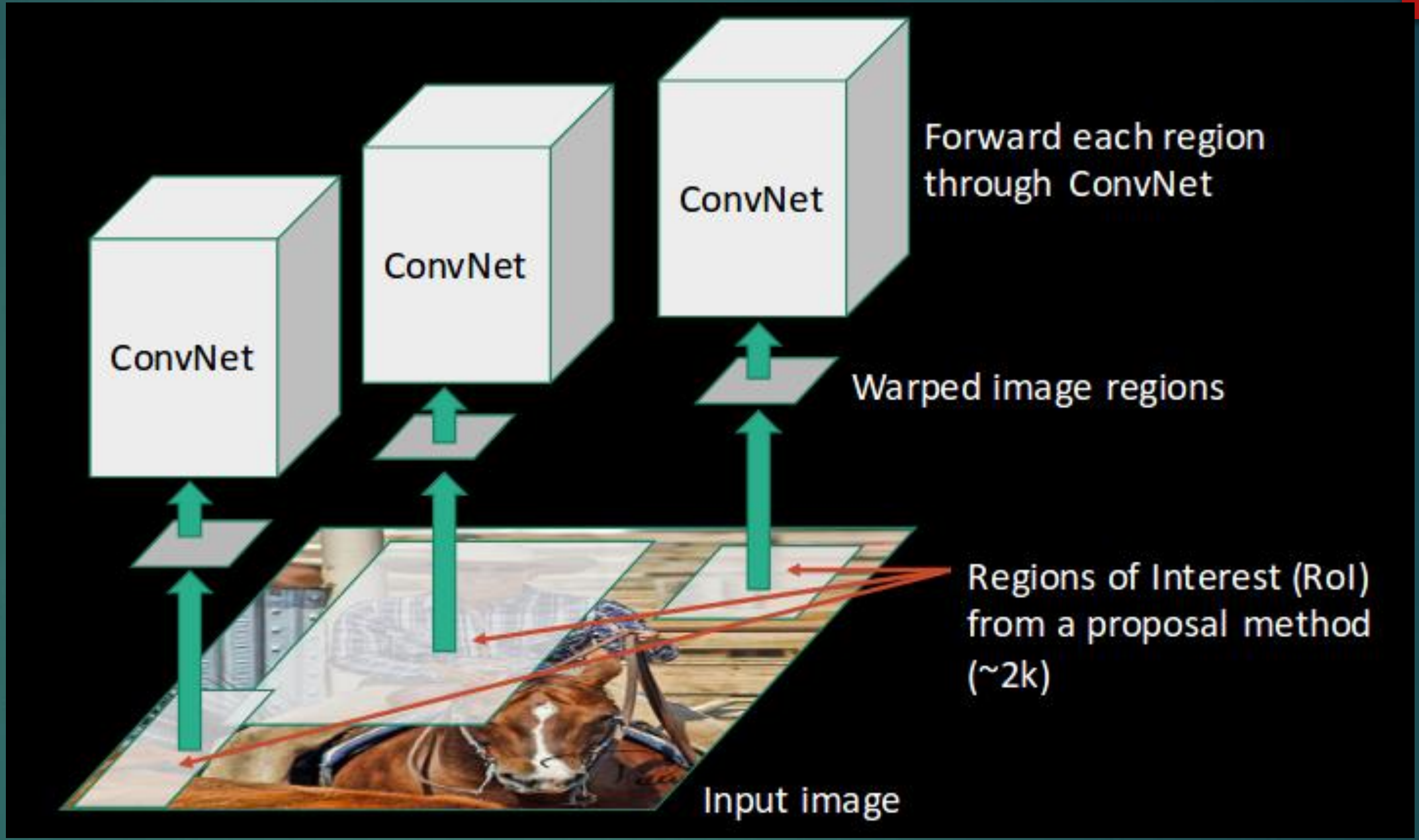
R-CNN Architecture

- First, an image is taken as an input:

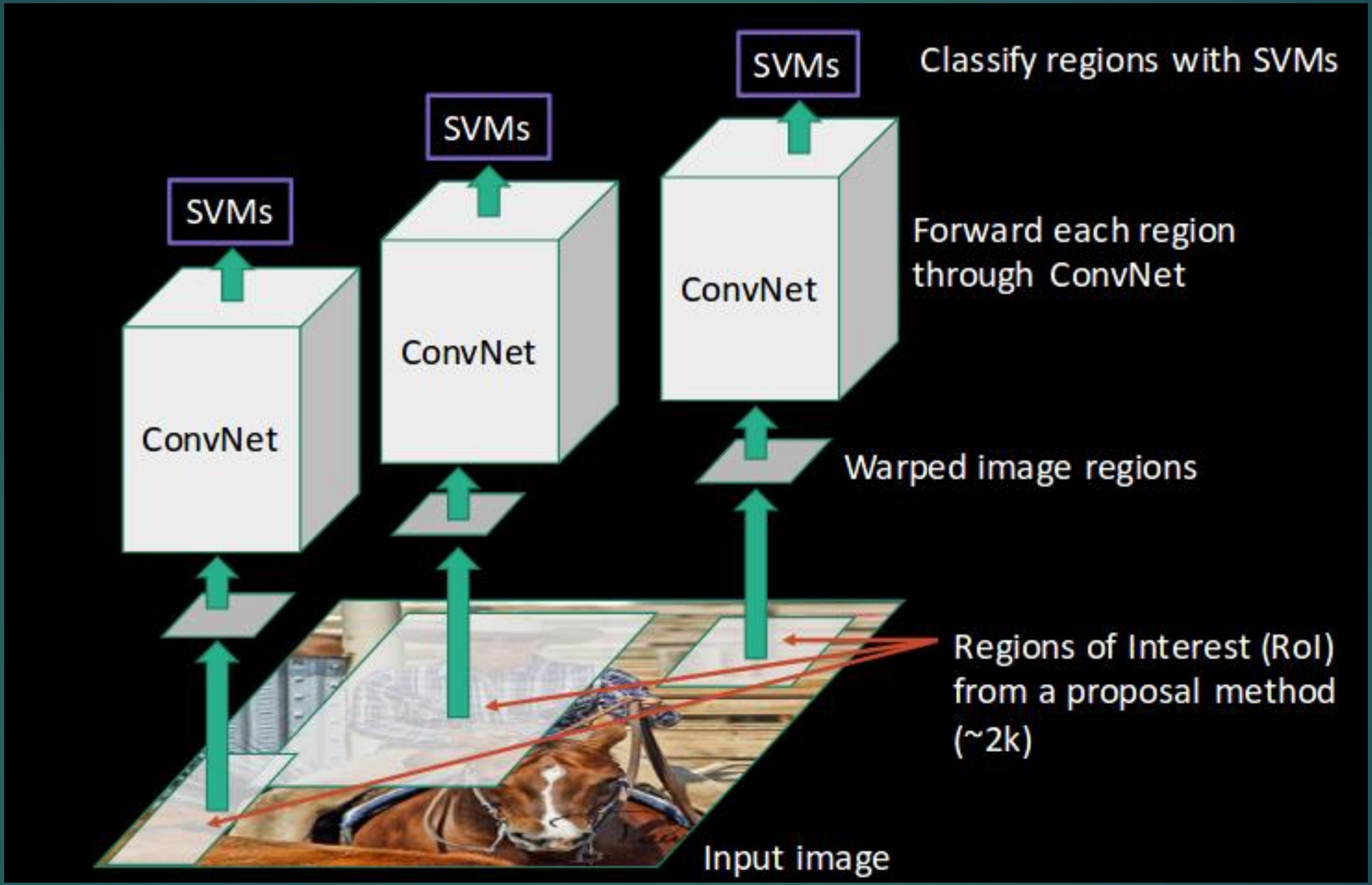
- Then, we get the Regions of Interest (ROI) using some proposal method



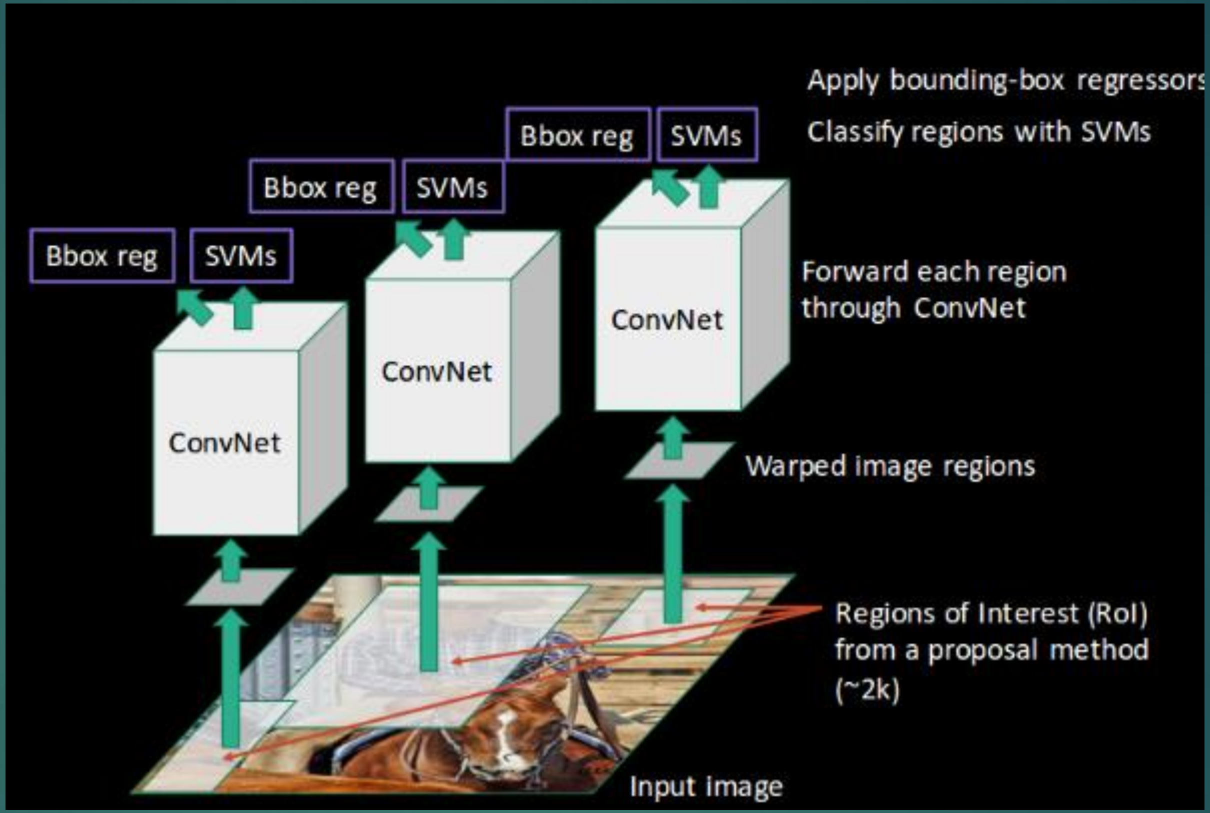
- All these regions are then reshaped as per the input of the CNN, and each region is passed to the ConvNet:


ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

- CNN then extracts features for each region and SVMs are used to divide these regions into different classes:


SVMs
SVMs
SVMs
Classify regions with SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

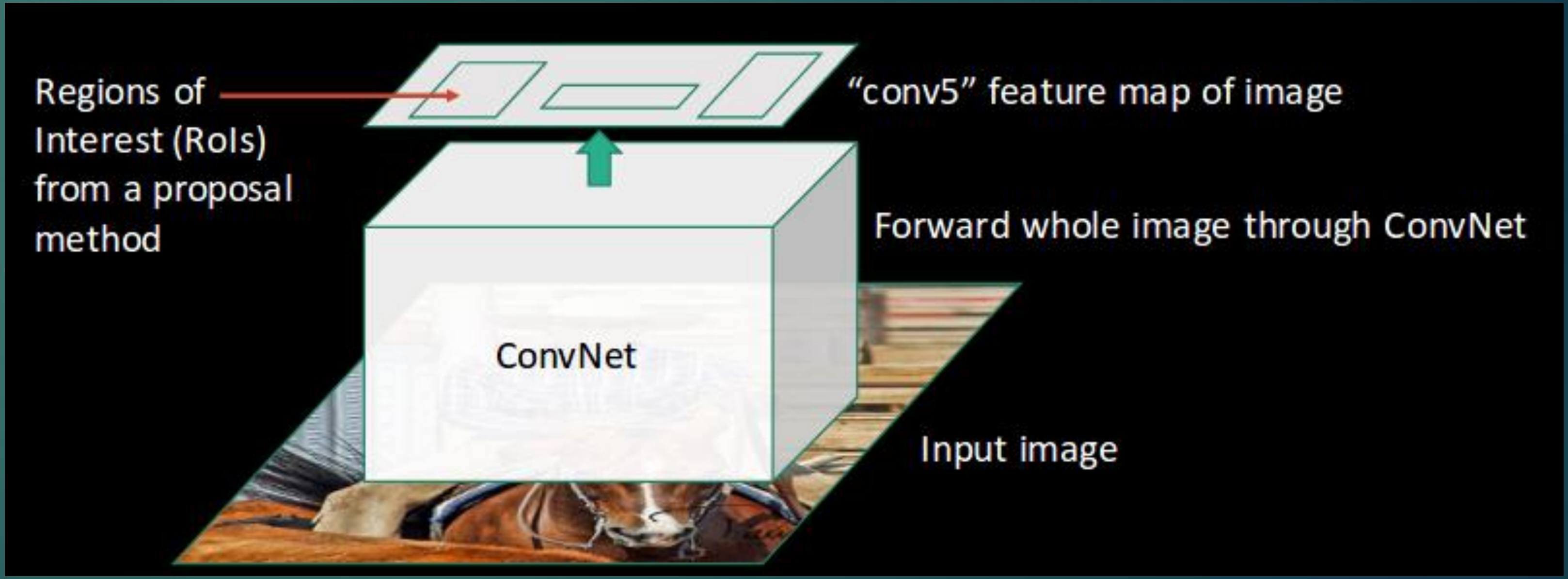
- Finally, a bounding box regression (*Bbox reg*) is used to predict the bounding boxes for each identified region:


Apply bounding-box regressors
Classify regions with SVMs
Bbox reg
SVMs
Bbox reg
SVMs
Bbox reg
SVMs
ConvNet
ConvNet
ConvNet
Forward each region through ConvNet
Warped image regions
Regions of Interest (RoI) from a proposal method (~2k)
Input image

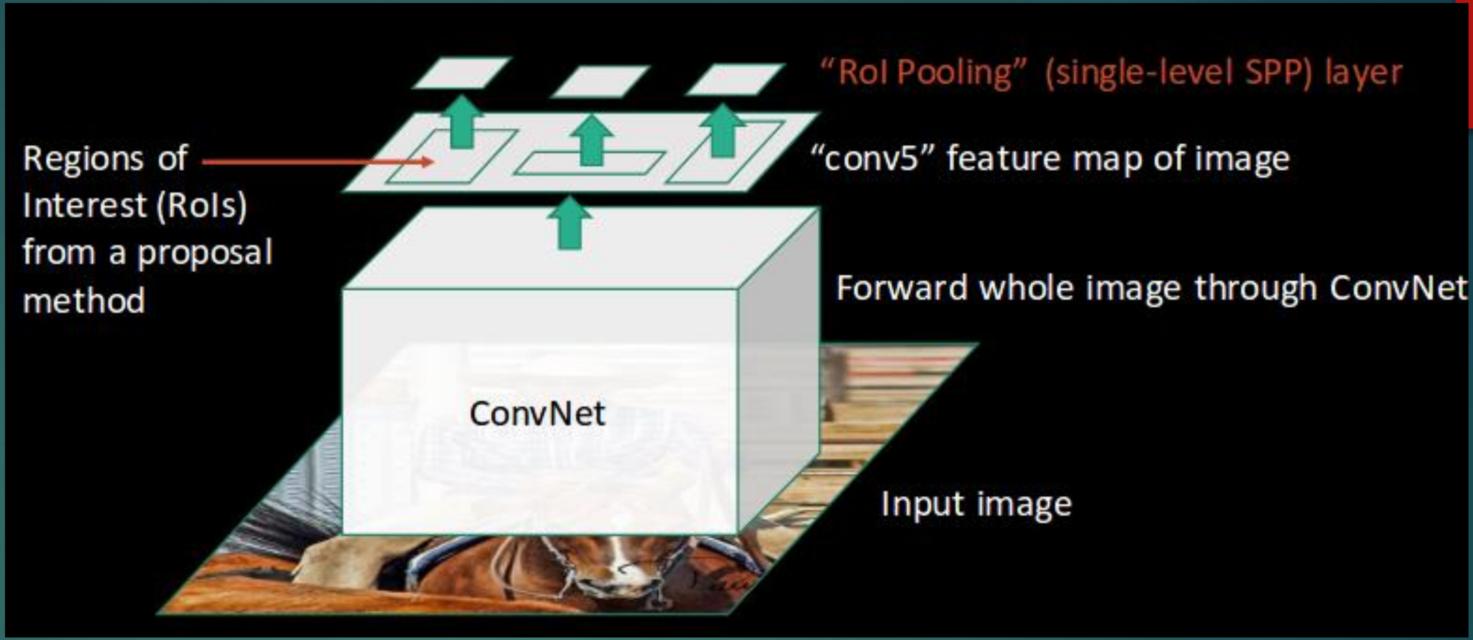
Fast R-CNN Architecture

- Input an image

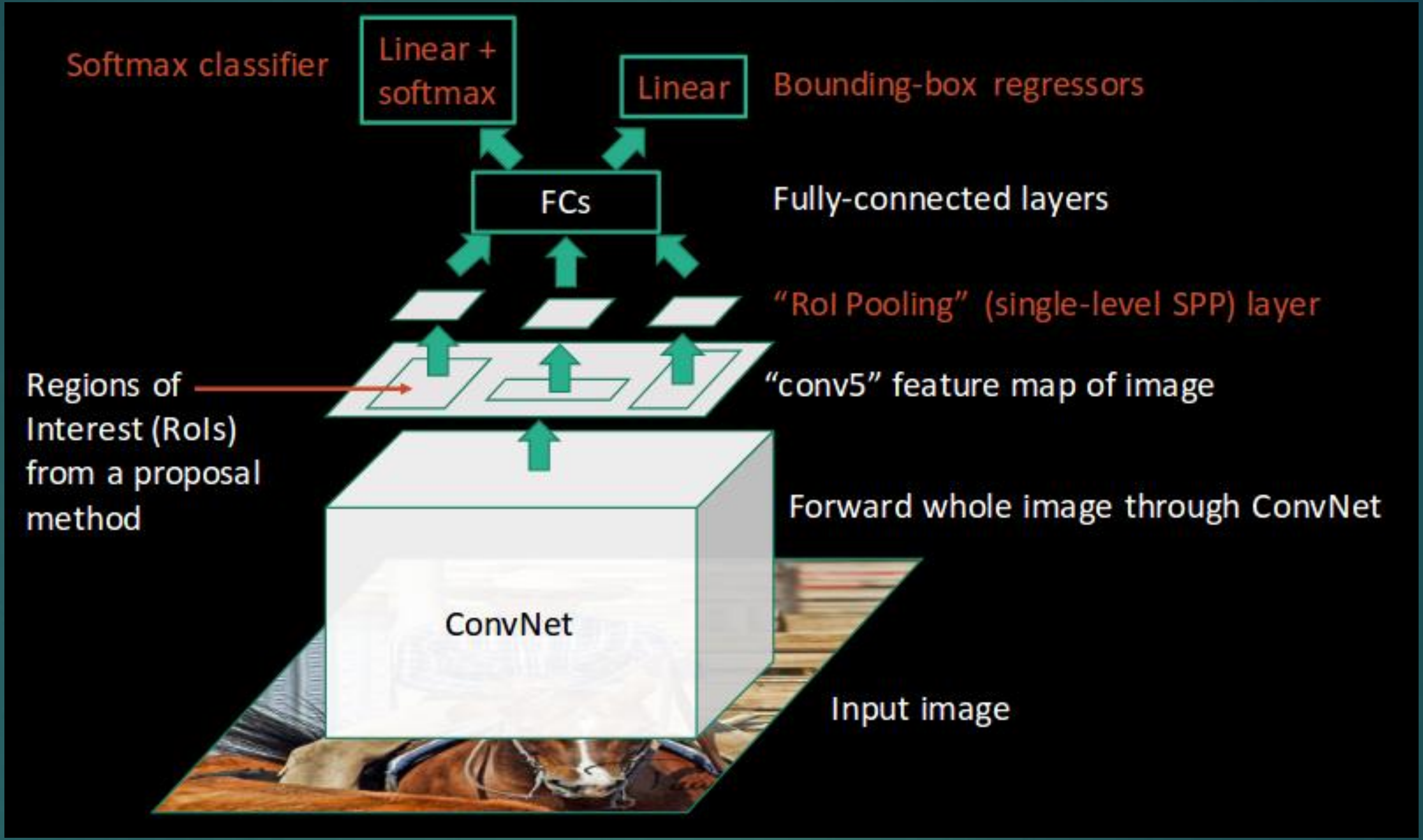
- This image is passed to a ConvNet which returns the region of interests accordingly:


Regions of Interest (RoIs) from a proposal method
"conv5" feature map of image
Forward whole image through ConvNet
ConvNet
Input image

- Then we apply the RoI pooling layer on the extracted regions of interest to make sure all the regions are of the same size:


"RoI Pooling" (single-level SPP) layer
Regions of Interest (RoIs) from a proposal method
"conv5" feature map of image
Forward whole image through ConvNet
ConvNet
Input image

- Finally, these regions are passed on to a fully connected network which classifies them, as well as returns the bounding boxes using softmax and linear regression layers simultaneously:


Softmax classifier
Linear + softmax
Linear
Bounding-box regressors
FCs
Fully-connected layers
"RoI Pooling" (single-level SPP) layer
Regions of Interest (RoIs) from a proposal method
"conv5" feature map of image
Forward whole image through ConvNet
ConvNet
Input image

Faster R-CNN Architecture

The below steps are typically followed in a Faster RCNN approach:

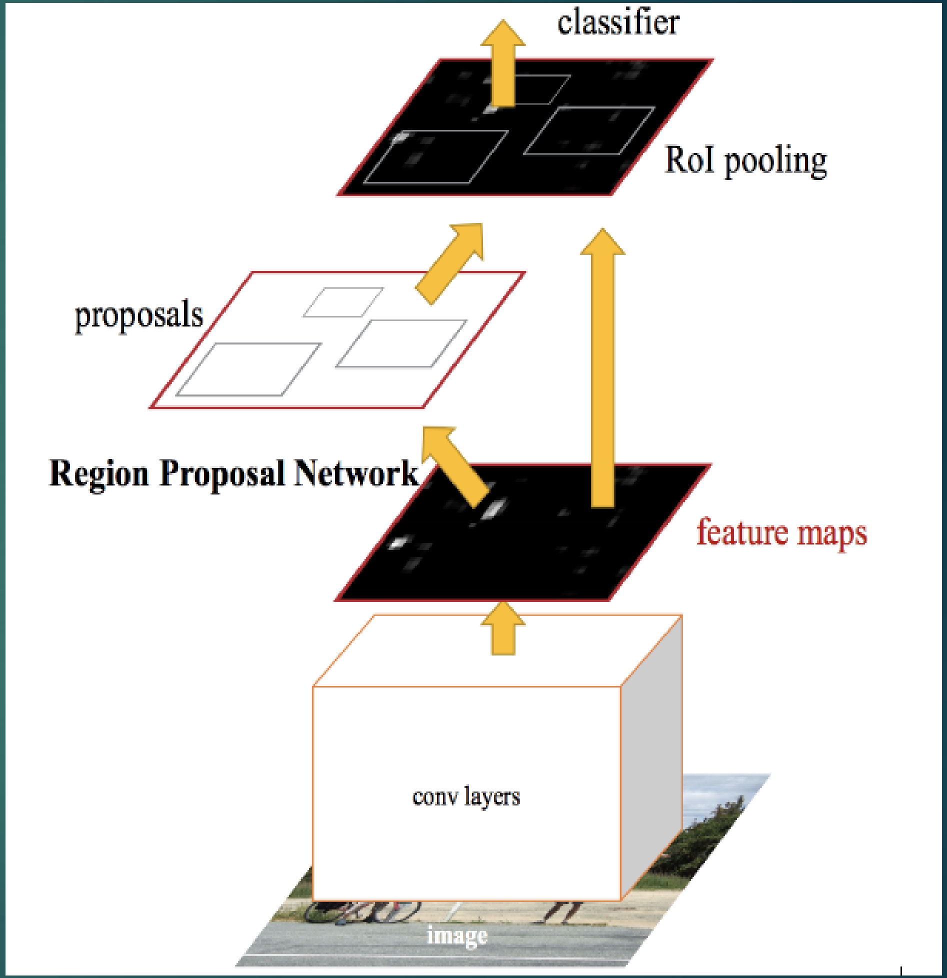
1. We take an image as input and pass it to the ConvNet which returns the feature map for that image.

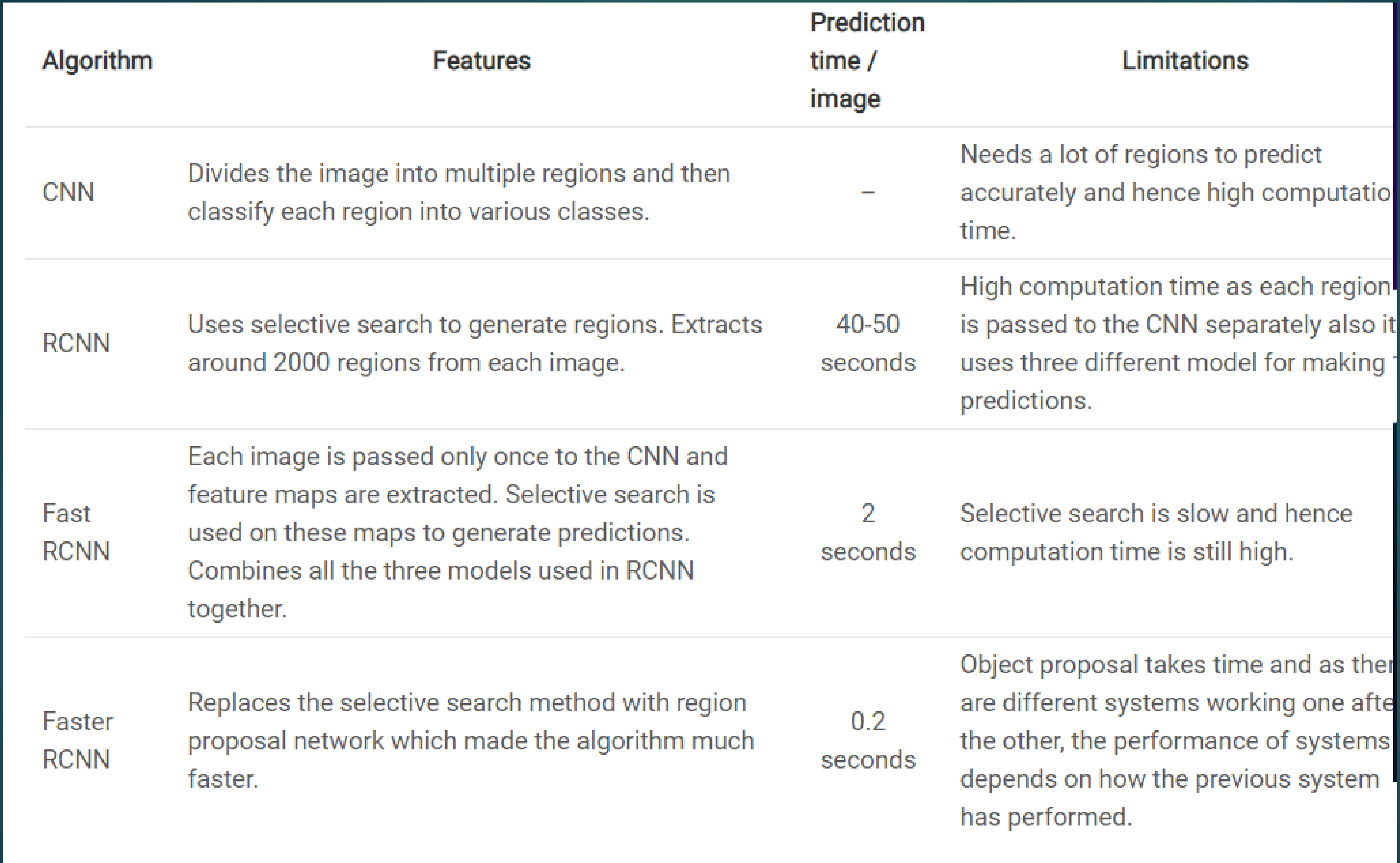
2. Region proposal network is applied on these feature maps.

This returns the object proposals along with their objectness score.

3. A RoI pooling layer is applied on these proposals to bring down all the proposals to the same size.

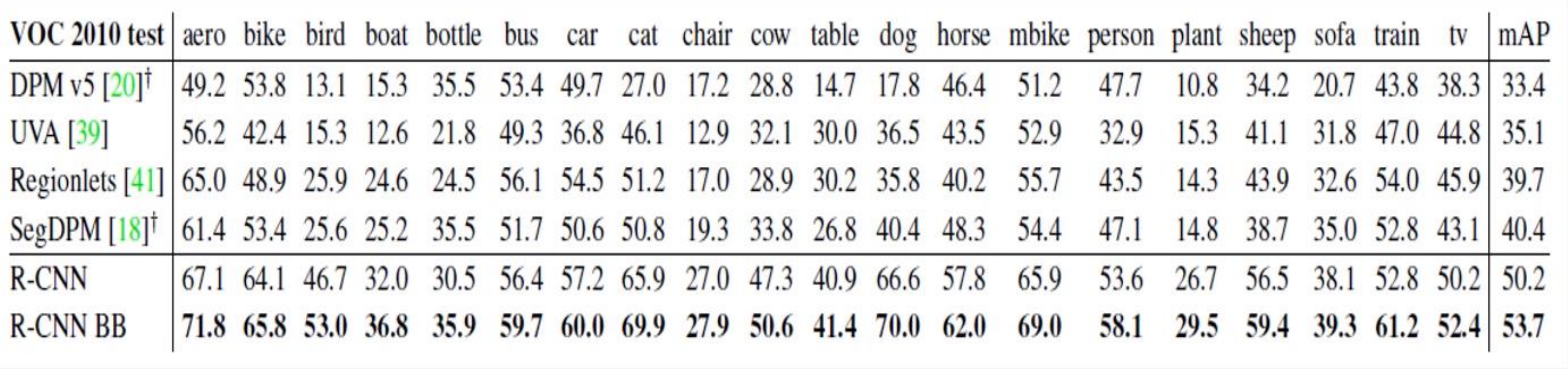
4. Finally, the proposals are passed to a fully connected layer which has a softmax layer and a linear regression layer at its top, to classify and output the bounding boxes for objects.





Algorithm	Features	Prediction time / image	Limitations
CNN	Divides the image into multiple regions and then classify each region into various classes.	–	Needs a lot of regions to predict accurately and hence high computatio time.
RCNN	Uses selective search to generate regions. Extracts around 2000 regions from each image.	40-50 seconds	High computation time as each region is passed to the CNN separately also it uses three different model for making predictions.
Fast RCNN	Each image is passed only once to the CNN and feature maps are extracted. Selective search is used on these maps to generate predictions. Combines all the three models used in RCNN together.	2 seconds	Selective search is slow and hence computation time is still high.
Faster RCNN	Replaces the selective search method with region proposal network which made the algorithm much faster.	0.2 seconds	Object proposal takes time and as ther are different systems working one afte the other, the performance of systems depends on how the previous system has performed.

Detection average precision (%) on VOC 2010 test



VOC 2010 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM v5 [20]†	49.2	53.8	13.1	15.3	35.5	53.4	49.7	27.0	17.2	28.8	14.7	17.8	46.4	51.2	47.7	10.8	34.2	20.7	43.8	38.3	33.4
UVA [39]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Regionlets [41]	65.0	48.9	25.9	24.6	24.5	56.1	54.5	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	43.9	32.6	54.0	45.9	39.7
SegDPM [18]†	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4
R-CNN	67.1	64.1	46.7	32.0	30.5	56.4	57.2	65.9	27.0	47.3	40.9	66.6	57.8	65.9	53.6	26.7	56.5	38.1	52.8	50.2	50.2
R-CNN BB	**71.8**	**65.8**	**53.0**	**36.8**	**35.9**	**59.7**	**60.0**	**69.9**	**27.9**	**50.6**	**41.4**	**70.0**	**62.0**	**69.0**	**58.1**	**29.5**	**59.4**	**39.3**	**61.2**	**52.4**	**53.7**

9/15/2020

Fig : Visual Object Classes benchmark 2010

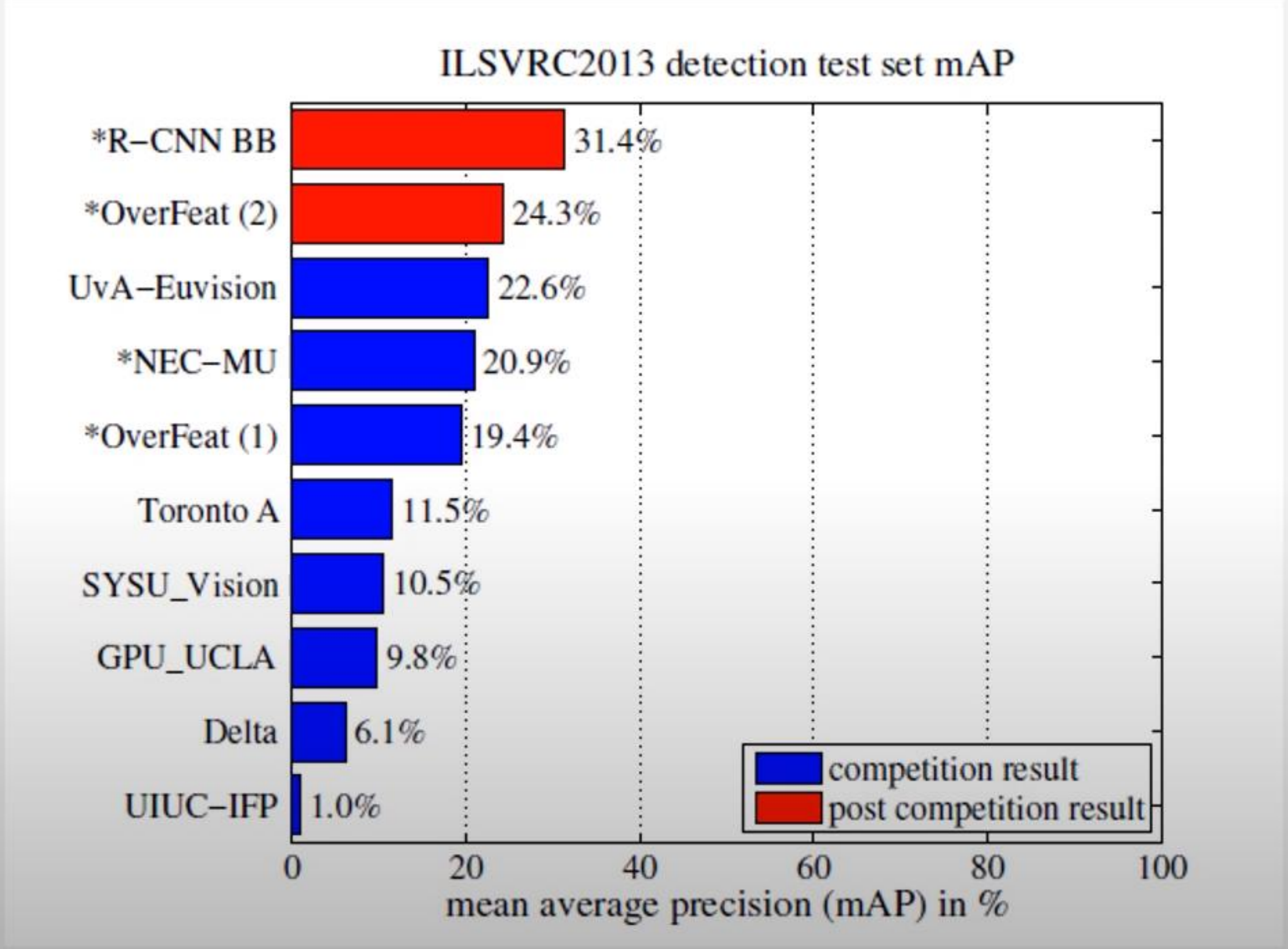
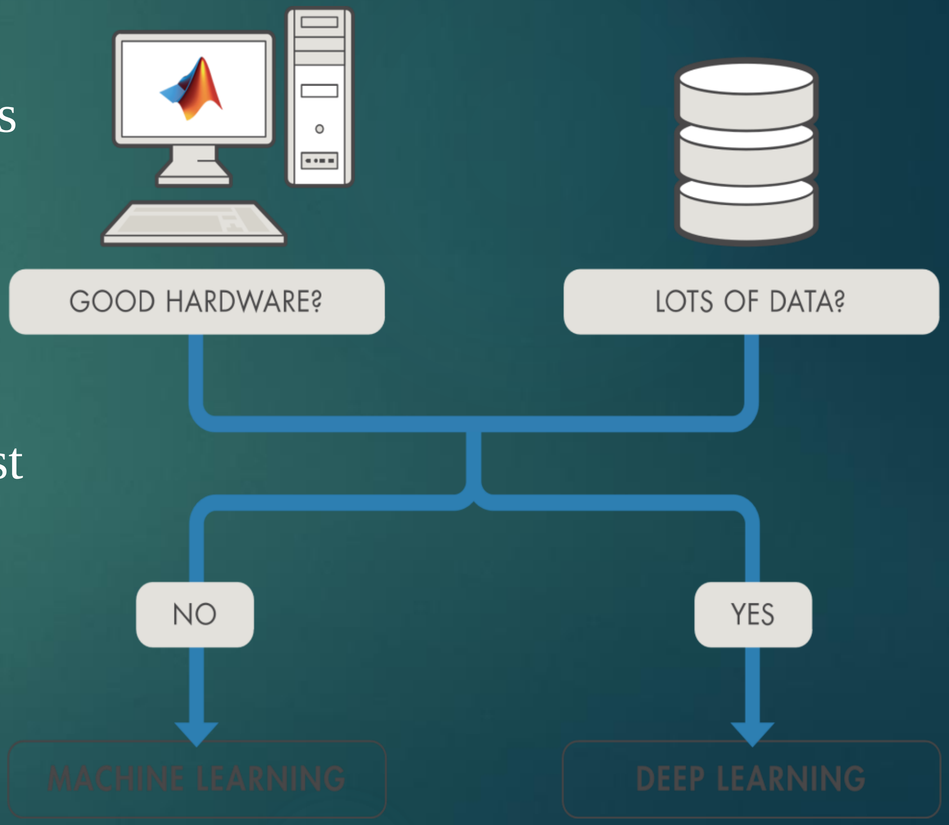


Fig : ImageNet Large Scale Visual Recognition Challenge benchmark 2013

Machine Learning vs Deep Learning for Object Recognition

- Machine learning for object recognition offers the flexibility to choose the best combination of features and classifiers for learning. It can achieve accurate results with minimal data.
- Determining the best approach for object recognition depends on our application and the problem we’re trying to solve. In many cases, machine learning can be an effective technique, especially if we know which features or characteristics of the image are the best ones to use to differentiate classes of objects.
- The main consideration to keep in mind when choosing between machine learning and deep learning is whether we have a powerful GPU and lots of labeled training images. If the answer to either of these questions is No, a machine learning approach might be the best choice. Deep learning techniques tend to work better with more images, and a GPU helps to decrease the time needed to train the model.



GOOD HARDWARE?

LOTS OF DATA?

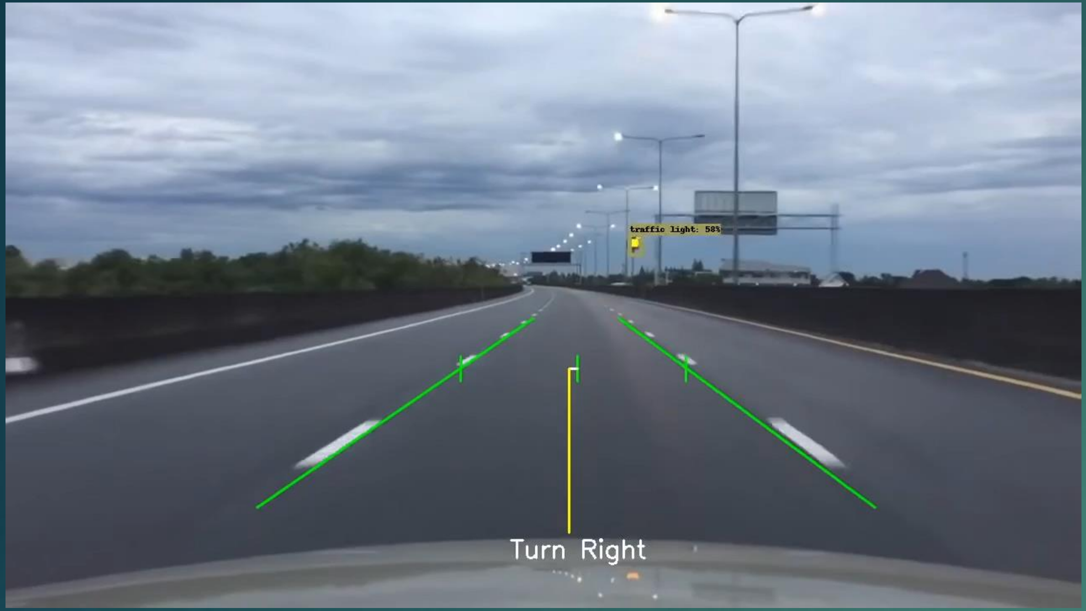
NO

YES

MACHINE LEARNING

DEEP LEARNING

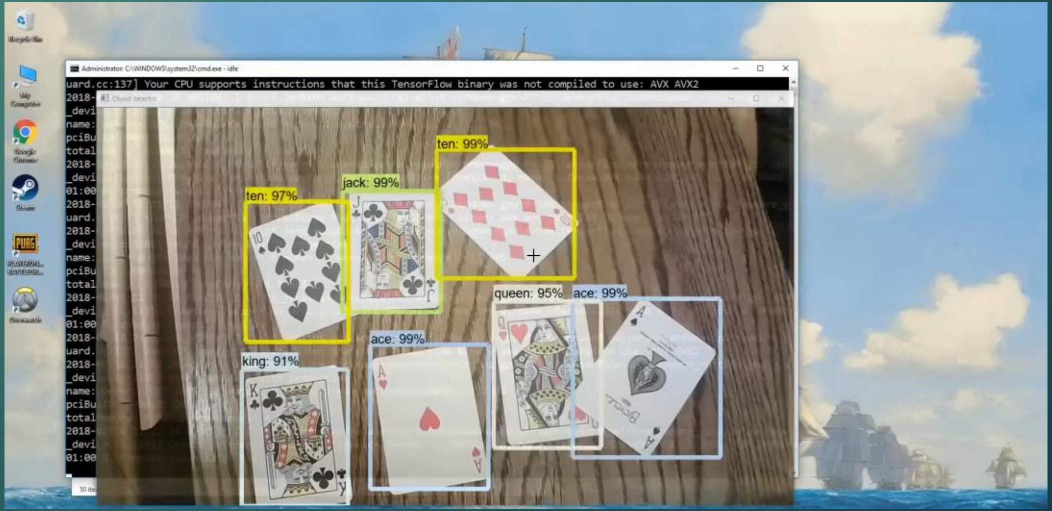
Some Application



Self -Driving Car

Surveillance Video

Crowd Counting



```
uard.cc:137] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX AVX2
```

Card Recognizer

Reference

- https://machinelearningmastery.com/object-recognition-with-deep-learning/
- https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/
- https://www.mathworks.com/solutions/image-video-processing/object-recognition.html#:~:text=Object%20recognition%20is%20a%20computer,%2C%20scenes%2C%20and%20visual%20details.
- https://www.fritz.ai/object-detection/

THANK YOU !