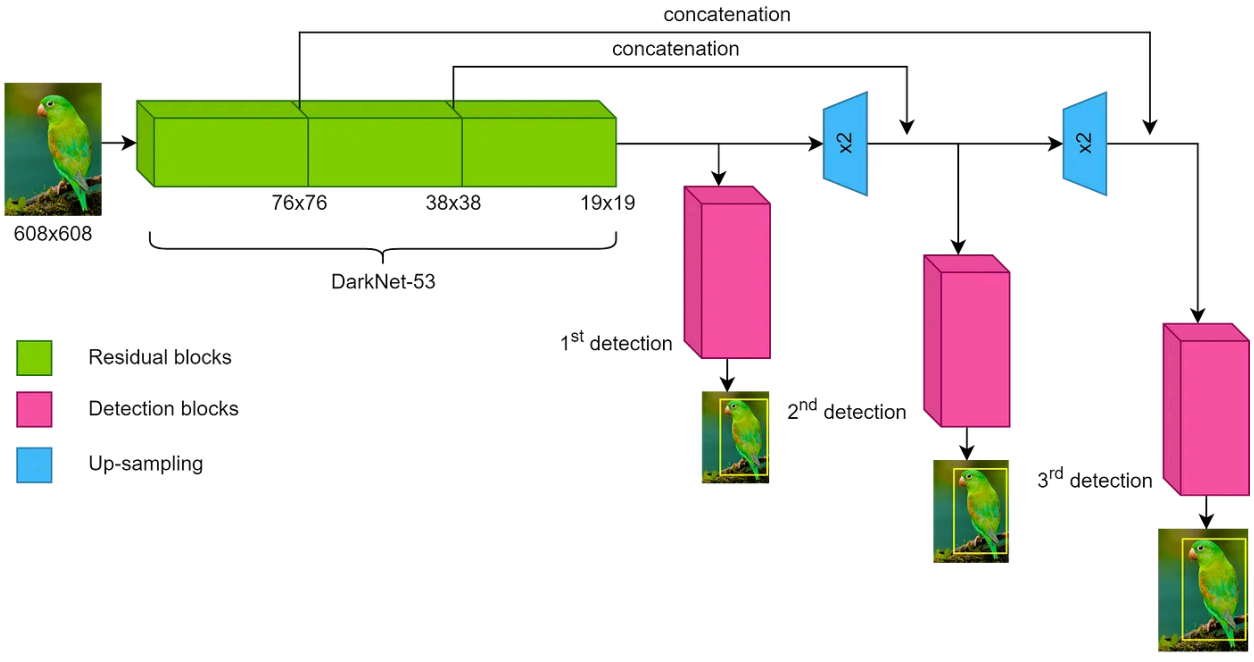
Report on Dynamic Object Recognition using YOLO v3

YOLOv3

YOLOv3, also known as 'You Only Look Once Version 3', is a real-time object detection algorithm that identifies specific objects in videos, live feeds or images. The YOLO machine learning algorithm uses features learned by a Deep Convolutional Neural Network to detect objects located in an image.

As typical for object detectors, the features learned by the convolutional layers are passed onto a classifier which makes the detection prediction. In YOLO, the prediction is based on a convolutional layer that uses 1x1 convolutions. YOLO stands for "you only look once" because its prediction uses 1x1 convolutions. This means that the size of the prediction map is exactly the size of the feature map before it.

How does YOLOv3 work??

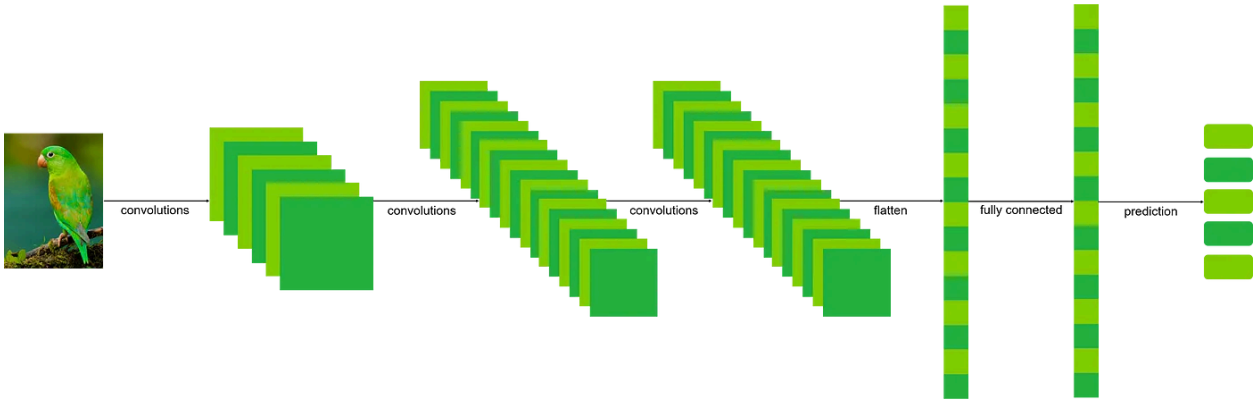


YOLO v3 architecture

YOLO is a convolutional neural network (CNN), a type of deep neural network, for performing object detection in real-time. CNNs are classifier based systems that process input images as structures arrays of data and recognise patterns between them. YOLO has the advantage of being much faster than other networks and still maintain accuracy.

CNN

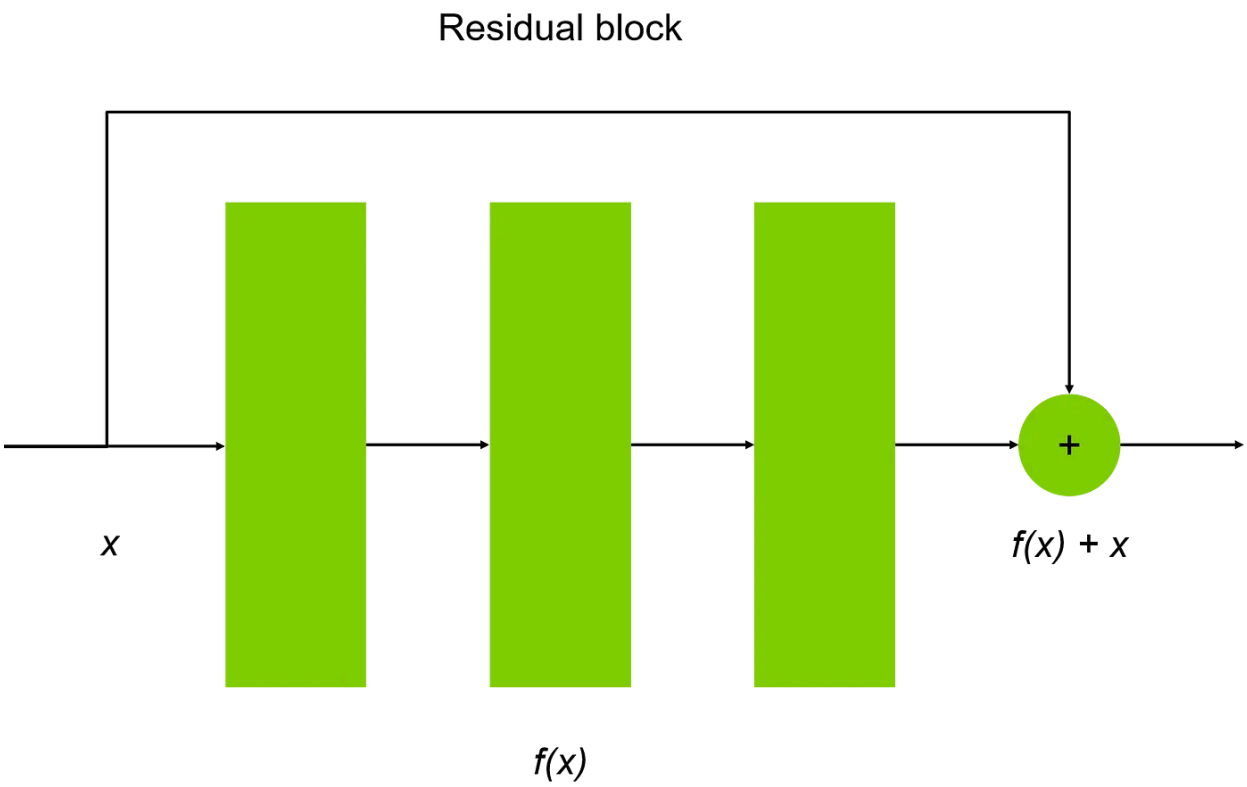
YOLO v3 is a CNN, a deep neural network specialized for image processing. Unlike traditional neural networks, it uses convolutional layers that apply learned filters (e.g., Gaussian blur, edge detection) to extract features. These filters, represented by kernels (e.g., 3x3 or 1x1 in YOLO v3), slide over the image with a specific stride, shaping the filter output dimensions.



Structure of a CNN

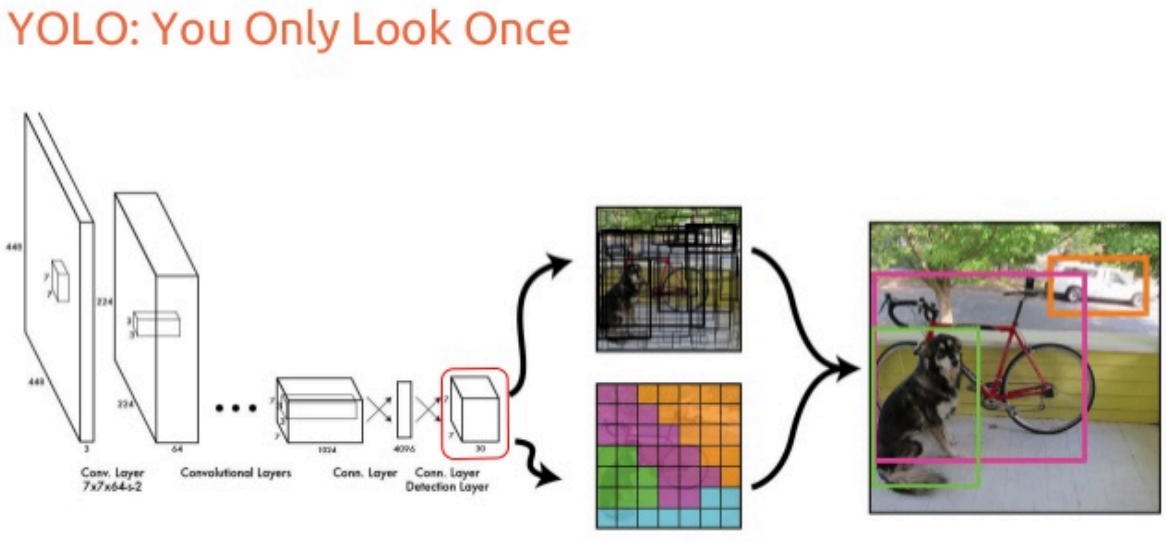
DarkNet-53

DarkNet-53, the YOLO v3 backbone, is a 53-layer CNN with 52 convolutional layers and skip connections. These connections address the challenge of deeper networks struggling to learn simple functions like identity mappings. Instead of passing output only to the next layer, skip connections allow it to bypass intermediate layers and be reused later. This approach separates simple information (residual block) from complex features extracted through deeper layers, combining them via summation for better learning efficiency.

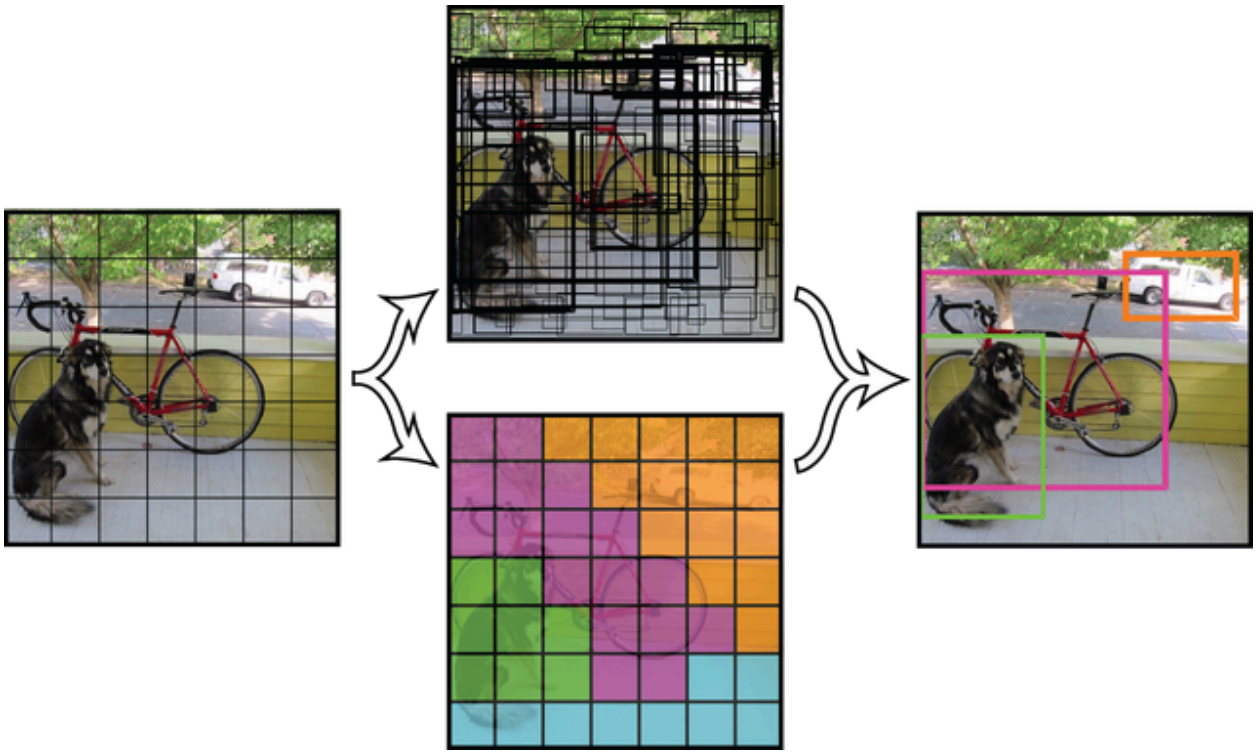


Structure of a residual block

Bounding Boxes



Anchor boxes are another key point of YOLO v3. They represent ideal bounding boxes and are discovered during the training phase by performing clustering on the known boxes of the examples. Therefore, anchor boxes describe the most recurring width/height ratios in the ground truth. For each scaling step (76x76, 38x38 and 19x19), three anchor boxes are extracted, for an overall of nine boxes. Anchor boxes are used at inference time to guide the definition of candidate bounding box size. In particular, bounding box size is expressed in terms of width and height ratio of the most suitable anchor box. Indeed, it has been observed that it is easier to express the size of a bounding box by considering anchor boxes, rather than starting completely from scratch.



Advantages of YOLO

1. Real-time speed
2. Fast processing suitable for video applications
3. Single network
4. Combines classification and localization in one pass
5. High accuracy
6. Effective in detecting multiple objects
7. Versatile

8. Works well in diverse environments
9. End-to-End Training
10. Streamlined training process
11. Generalisation
12. Robust across various contexts and objects

Disadvantages of YOLO

1. **Smaller Objects Detection:** YOLOv3 can struggle to detect smaller objects due to its anchor box design and large stride.
2. **High Memory Requirements:** YOLOv3 requires a significant amount of memory to run, which can be a challenge for devices with limited resources.
3. **Training Time:** Training YOLOv3 can be time-consuming, requiring large amounts of data and computational resources.

References

1. https://medium.com/data-reply-it-datatech/yolo-v3-for-object-detection-f9f5637cf428
2. https://medium.com/analytics-vidhya/understanding-yolo-and-implementing-yolo v3-for-object-detection-5f1f748cc63a