Machine Learning for Object Detection

By Ashish Bhattarai

November 15, 2021

Introduction

Object detection is one of the widely researched topics inside computer vision. To define it in simple terms, it is the task of identifying the regions containing relevant objects from rest of the scene inside a digital image. In object detection, the problem is usually to locate approximate bounding boxes around the objects with minimal region from background or other objects.

The applications for object detection capability can be wide-ranging. For instance, one could locate a person, a chair or a dog inside an image of a room, or cars, traffic lights, pedestrians inside image from a road, or locate cells and identify their phenotype inside medical images, or find texts inside photos, or get facial region inside a photo for facial recognition task, and so on.

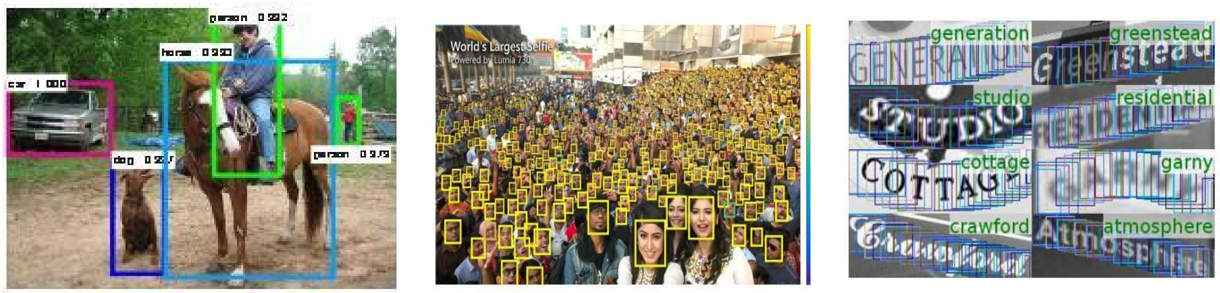

Figure 1: Detection of different objects in scenes [1-3]

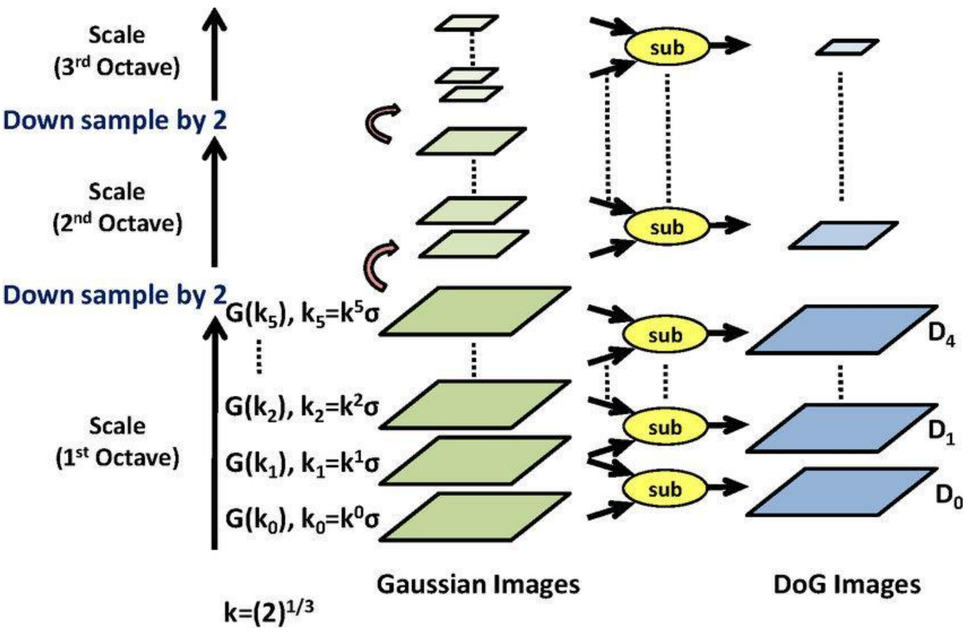

Figure 2: Gaussian (**left**) and DoG (**right**) feature pyramids [4, 5]

How Does It Work?

Let's first consider at an intuitive level how one would go about solving this problem of locating objects. To form a more proper idea, let's first consider what properties characterize certain objects inside a scene, or to be more particular, let's consider what constitutes a person in the eye of us humans. Description of clothes can vary from person to person, so cannot be a reliable indicator. Tone of skin changes but has more limited variations and can be slightly helpful. Beyond such color information, perhaps it's the shape of a person, the contours that is more relevant. Visual perception of distinct body parts noticeable at different scales: head, hands, legs, torso at coarse level to small parts such as eyes, ears, fingers, and so on at finer levels. Not all of these may be simultaneously visible, but they provide a basis for fairly unique and reliable features to distinguish a person from, say, a chair, a tree, or a car.

Digital representation of a 2D real-world image is just 2D grid of numbers. We require tools from mathematics to deal with the numbers. Through millions of years of evolution from single cell organisms to large beings with complex body parts, we humans have developed great ability to interact with our surroundings in effortless manner. But a good mathematical translation of this skill, in particular, the power of vision for identifying objects inside digital images is non-trivial task.

Traditional Methods

For object detection, popular traditional approaches relied on matching of multi-scale representations (see Fig. 2) of some reference template models to detect the objects in the scene. The use of multi-level representations provides scale invariance, which has been observed in behavior of complex cells in the cerebral cortex of mammalian vision. This concept underlies some of the well-known classical approaches such as Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) object detector. Aside from these, Haar feature matching was used in Viola-Jones detector for face detection, whereas sliding window-based matching was popularized through works of Histogram of Oriented Gradients (HOG) detector and Deformable Parts Model (DPM).

The feature descriptors could be changed to some other representation, such as Gradient Location and Orientation Histogram (GLOH), Local Energy based Shape Histogram (LESH), or an ensemble, depending on the suitability to the problem. Such features covered solely the explicitly derivable properties such as color values, gradients, hessians, wavelet features, or specific texture properties.

Use of explicit features can be reliable but requires domain-expertise and careful engineering. Further difficulties with these approaches are that they are slow, inaccurate and lack good generalizability to a large dataset.

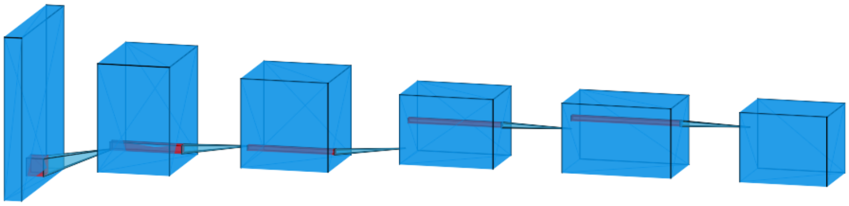

Figure 3: Multiscale architecture of typical CNN based feature extractor

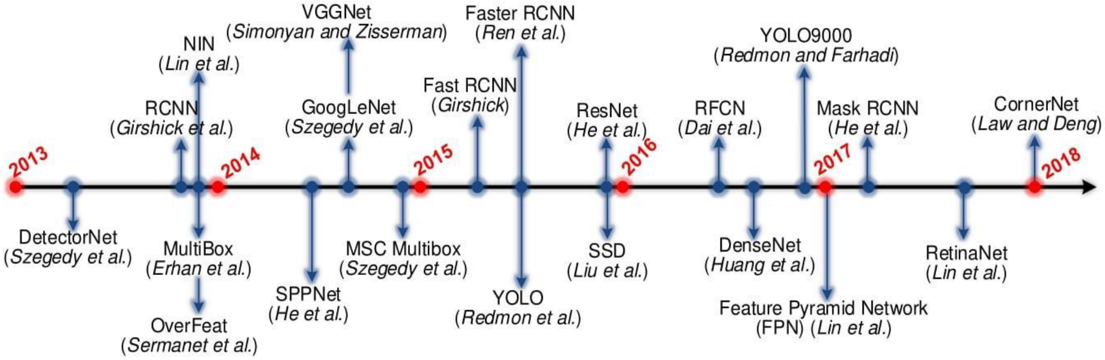

Figure 4: Milestones in deep learning based object detection [6]

Deep Learning Based Techniques

The paradigm shift with deep learning based approaches has been in the transition from the explicit to the implicit feature representations. The state-of-the-art object detection methods today are neural networks based. Several benchmarks such as PASCAL VOC, MS COCO, KITTI, ImageNet are frequently used for performance comparisons of different learning based models on object detection architectures. The number of object categories is predefined. For instance, there are 80 objects in MS COCO, 20 in PASCAL VOC, 200 ILSVRC challenge. Widely used metric for evaluation of results is the average precision (AP) for individual category and mean average precision (mAP) over all categories.

The initial convolutional neural network (CNN) based object detection models such as Region based CNN (RCNN) and Faster RCNN were based on the idea of sliding window based matching similar to the classical methods, but instead going over a 2D grid of transformed image features. However, nowadays modern CNN (CNN) based object detectors are typically single stage. This has been achieved by treating object detection as a multi-value regression task at each pixel and minimizing the distance between predicted and ground bounding box coordinates.

An obvious difficulty encountered while using single stage detectors like YOLO and SSD faced for regressing bounding box coordinates at a smaller grid size had been the detection of small objects. To improve results, more advanced models like YOLOv2, RetinaNet advocated combining the multi-scale features with short-cut connections, use of better backbone networks, batch normalization layers, joint classification and bounding box regression, and/or improved loss function that can handle class imbalance. Recent developments have considered constructing better feature pyramid representations with decoder blocks inside backbone network, using transformer based backbone and applying deformable convolution kernels to adapt to different structures of objects.

Though deep learning methods provide state of the art results in object detection tasks, large intra-class variance as well as different imaging conditions create extra challenge to reach a significantly high accuracy.

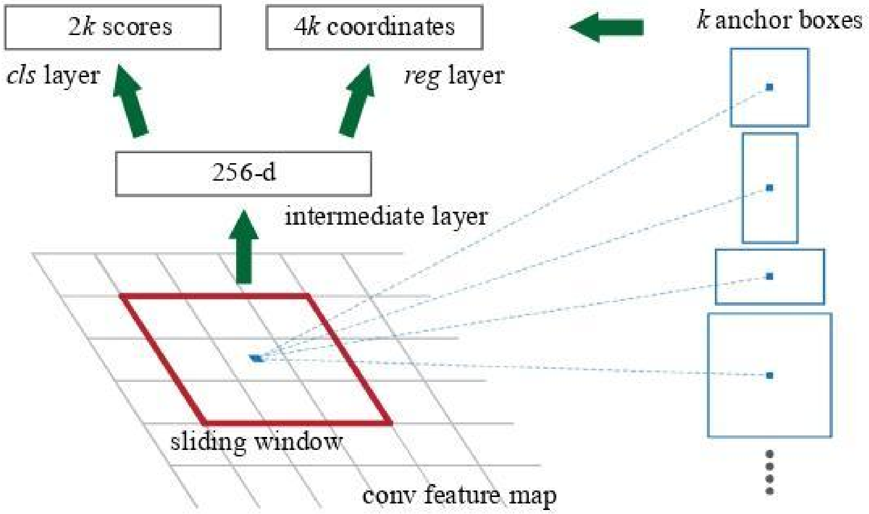

Figure 5: Sliding window based feature matching with Region Proposal Network from Faster-RCNN [1]

Final Comments

Even though classical non-learning-based approaches suffer from severe drawbacks in terms of performance and accuracy, they are still useful for very low dataset conditions and where explainability becomes critical, such as in medical image analysis tasks. Neural networks, on the other hand, are somewhat black-box constructions in that the final implicit feature space transforms can lack reasoning, and might simply be overfitting the dataset.

Object detection task should be distinguished from the other related task of image segmentation, which is instead a pixel-wise classification problem. Models like YOLOv2 use pixel-wise objectness score to improve results and one might wonder, why not always segment the objects before and locate them afterwards. Even though that seems to be a reasonable principle, the difficulty arises from the fact that precise semantic segmentation itself turns out to be more challenging. Furthermore, a mislabeled pixel during segmentation can result in severely imprecise bounding boxes. Therefore, in practical settings object detection task is treated separately from the semantic segmentation task.

For efficient utilization of the hardware and low latency inference results, a new research area in the realm of light-weight object detection models has developed. Some such architectures are the MobileNet, SqueezeDet, MnasNet, and light variants of different YOLO models.

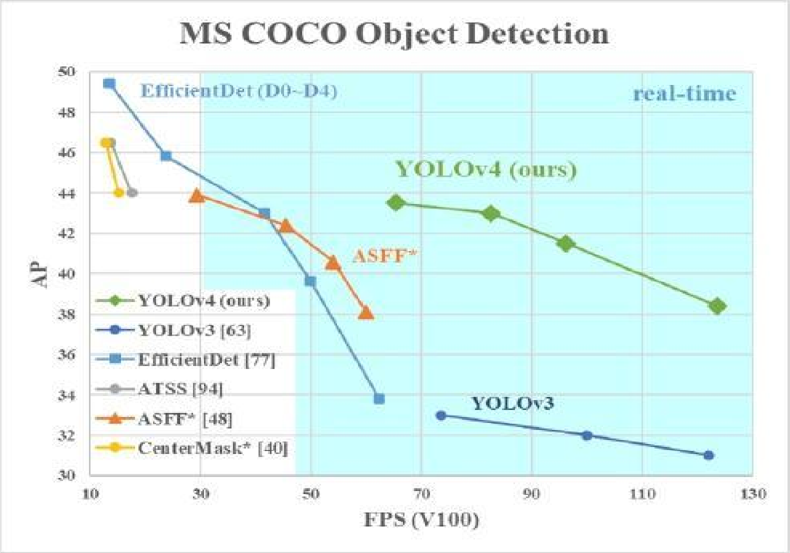

Figure 6: Performance comparison of object detection models with respect to YOLOv4 model [7]

References

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[2] P. Hu and D. Ramanan, “Finding tiny faces,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 951–959, 2017.

[3] Bartz C., Yang H., and Meinel C. STN-OCR: A single neural network for text detection and text recognition. CoRR, abs/1707.08831, 2017.

[4] Lowe D. (1999) Object recognition from local scale invariant features. In: ICCV, vol 2, pp. 1150–1157.

[5] Huang F., Huang S., Ker J. and Chen Y., "High-Performance SIFT Hardware Accelerator for Real-Time Image Feature Extraction," IEEE Trans. on Circuits and Syst. for Video Tech., vol. 22, no. 3, pp. 340-351, March 2012.

[6] Liu L., Ouyang W., Wang X., Fieguth P., Chen J., Liu X., and Pietikainen M., “Deep learning for generic object detection: A survey,” arXiv preprint arXiv:1809.02165, 2018.

[7] Bochkovskiy A., Wang C.-Y., and Liao H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.