PROJECT REPORT

COURSE: ADVANCED PROBLEMS IN COMPUTER SCIENCE

OBJECT DETECTION IN DEEP LEARNING

And their applications in real life

Dung Nguyen Manh

Academic year: 2020 - 2021

CONTENTS

Abstract

For decades, people are dreaming about create a machine with the characteristics of human intelligence, those can think and act like human. Nowadays, thanks to the advancements of artificial intelligence and computational power, Computer Vision technology has taken a big evolution and significant role in enabling digital transformation across different industry. Computer Vision technology is transformating the busniess world with its capability to understand the content of digital images and videos. Accouding to Tractica [1], global market for computer vision will increase from $6.6 billion in 2015 to $48.6 billion annually by 2022, which re-confirm the huge impact of Computer Vision fields to this world. With the concept of capturing, processing, analyzing digital images and videos, Computer Vision allows computer to see and understand the real world and generates actionable insights as per designed algorithms. In this report, I will cover the definition of Computer Vision and the most basic concept of its subfields, especial Object Detection.

1 INTRODUCTION TO COMPUTER VISION

Computer Vision (CV) is a subfields of Aritficial Intelligence (AI), emerged in the late 60's and developed almost parallely with the AI field. The term "Computer Vision" have 2 components, where "Computer" refers an electronic machine capable of performing various processes, calculation, and operations from sets of instructions directed by software or hardware, and the term "Vision" refers to visual perception throught sight where can be understand as the ability to "identify" the objects located inside the environments.

Figure 1: CV agent is an AI that can interpret and understand the visual world

Source: *dreamstime.com*

The concept of Computer Vision is based on teaching the machine how to “see” and interpret important information contained in images and videos. Computer Vision systems then use this translated data, using the knownledge provided by human beings, in order to improve the result of decision making process. Turning raw image data into higher-level concepts, that computers can interpret and act upon them is the principal goal of computer vision technology.

1.1 A brief history of Computer Vision

Computer Vision is not new technology; the first experiments with Computer Vision started in the summer of 1966, Seymour Papert and Marvin Minsky started a project titled "Summer Vision Project" [2], where they built a system that can analyse a scence and identify objects in that scence. At that time, computer vision were relatively simple and required a lot of work from human operators who had to provide data samples for analyse manually. It’s hard to provide a decent amount of data, plus, the computational power that day was not enough, therefore the error margin in this project was pretty high.

Figure 2: Seymour Papert and Marvin Minsky in The History of AI at AIWS.net

At first, low-level tasks such as color segmentation or edge detection, etc, were already applied in the early day of the fields and formed the foundations of many mordern computer vision this day. However, by the 80’s, the scientific world generally agreed that the problem was not as trivial as they initially thought it was. Scientists quickly came to realise that tasks that are easily or even unconsciously done by humans are very difficult for a computer and the opposite.

In late 70's and early 90's, knowns as "AI Winters", is a period of reduced funding and interest in artificial intelligence research [3]. A principle, commonly known as Moravec's paradox [4], was first formulated by the computer scientist Hans Moravec. Basically, it highlights that is much easier to implement specialized computers to mimic adult human experts (Deep Blue beat Kasparov at chess [5]) than building a machine with skills of 1-year old children with abilities to learn how to move around, recognize faces and voice or pay attention to interesting things.

> Easy problems are hard and require enormous computation resources, hard problems are easy and require very little computation. (Moravec's paradox [4])

The "winter" of connectionist research came to an end in the middle 1980s, when the work of John Hopfield, David Rumelhart and others revived large scale interest in neural networks. The mid 90's, the field has seen an increase in interest with the widespread use of machine learning and the first industrial applications. Scientists in Machine Learning started to shifts from a knowledge-driven approach to a data-driven approach, and many technical machine learning arrived such as Support-vector machines (SVM); Recurrent neural networks (RNN); etc.[6, 7] In the past decade, the introduction of deep learning has reinforced the interest in the field, intensifying the talk about an "AI spring".

1.2 Challenge in Computer Vision

The main purpose in development Computer Vision is not to imitate just human sight, but actually to imitate the human visual perception. When parallelize the human visual system with the computer vision system, we can say both consist of sensor and interpreter. If our eyes are our sensor that help us senses the light, then camera in many devices are machine 'eyes'. The biggest challenge comes up with Computer Vision is not in machine eyes but the algorithm teach them how to observate the world.

A computer sees an image as series of pixels with it transforms to and array. The figure 3 is a simple illustration of the grayscale image buffer which stores our image of Abraham Lincoln. Each pixel's brightness is represented by a single 8-bit number, whose range is from 0 (black) to 255 (white). Nowadays, most of the images have 3 color channels which is RGB, then each channels no more represents grayscale intensity but the related color intensity, eg: channels Red represents the intensity of color red in image.

Because of the way computer or machine seeing the world, Computer Vision have to face many challenges. Therefore, it's not easy to answer what is the challenges in Computer

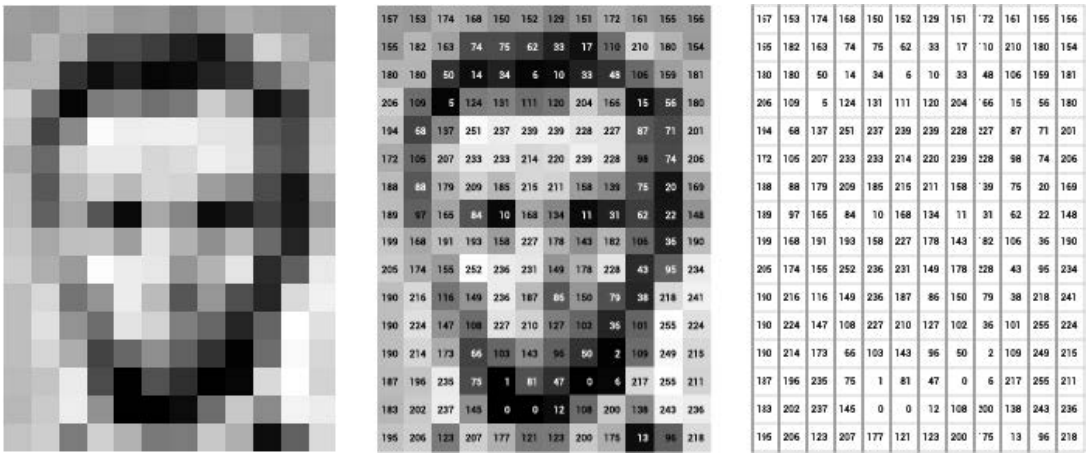

Figure 3: How computer interpret image

Source: *Openframeworks*

Vision, these difficulties might vary from each to other fields in Computer Vision. The mose comman challenges in CV tasks such as objects under different viewpoints, illuminations, and intraclass variations, etc.

Imagine we are building a Computer Vision system has ability to classify whether the objects in image are cat or not. That a look at a few example in figure 4 which can show us what type of difficulties that real life image bright to our model. It's clearly that with the same cat, but if we change the camera angle, the system will definitely get another image which is unrelative to any other image that the system has seen.

Figure 4: Challenge on first day in Computer Vision

Source: *CS231n Stanford*

2 WHAT IS OBJECT DETECTION

We can not deny that Computer Vision, at the moments, take an important role in many applications. It takes part in many intelligence system that are able to understand and finish their task through a process of interpreting the visual world by camera, videos, images, etc. As one of the fundamental problems of computer vision, Object Detection form the basic of many other computer vision task, such as instance segmentation, image captioning, object tracking, etc.

Object detection (OD) is a computer technology related image processing that deals with detecting instances of semantic objects of a certain class (such as humans, animals, or cars) in digital images and videos. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: *What objects are where?*

While Object detection is a subfiled of computer vision, OD also face the most basic difficulties in computer vision that we covered in section 1.2. However, object detection include (but not limited to) the following aspects: object rotation and scale changes, accurate object localization, dense and occluded object detection, seed up of detection, etc.
Under the revolution of Deep learning, object detection has achieved many accomplishments. Deep learning has lead object detection to remarkable breakthourgh and pushing it forward to a research hot-spot with unprecedented attention. The number of publications in object detection from 1998 to 2018 has increasing rapidly. [8]

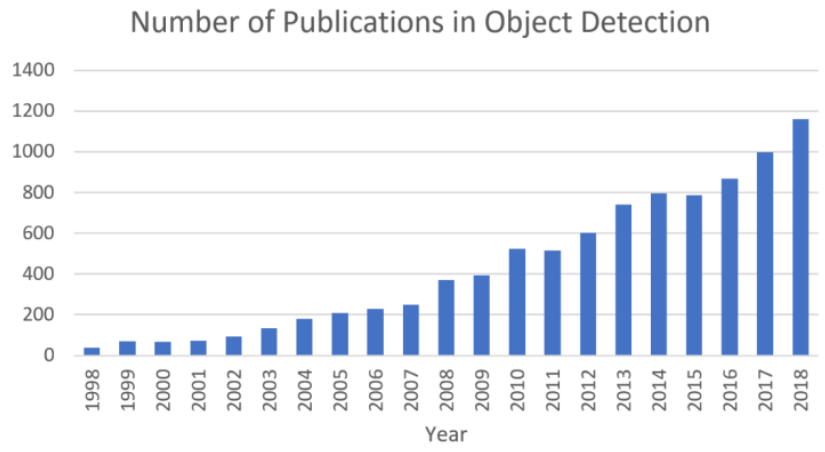

Figure 5: The increasing number of publications in object detection from 1998 to 2018

Source: *Object Detection in 20 Years: A survey [8]*

3 A ROAD MAP OF OBJECT DETECTION

In the past two decades, the scientists widely argeed that object detection has gone through two historical periods: "traditional object detection" (before 2014) and "deep learning based detection" (after 2014).

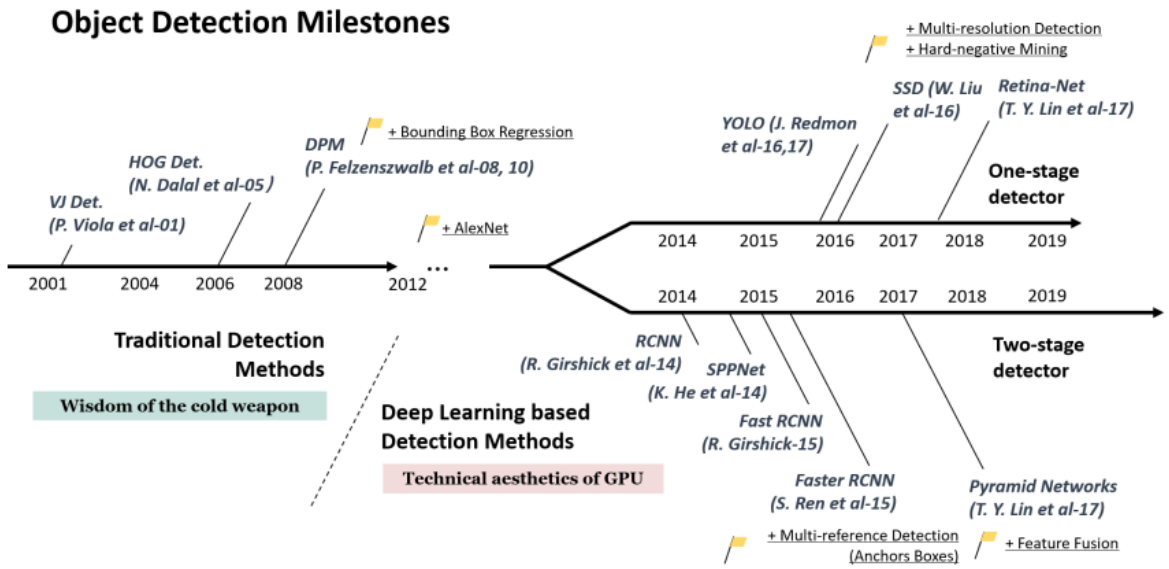

Figure 6: A road map of object detection. Milestone detectors in this figure: VJ Det. [9, 10], HOG Det. [11], DPM [12, 13, 14], RCNN [15], SPPNet [16], Fast RCNN [17], Faster RCNN [18], YOLO [19], SSD [20], Pyramid Networks [21], Retina-Net [22].

Source: *Object Detection in 20 Years: A survey [8]*

3.1 Traditional Object Detectors

Turn back the clock 20 years, we would witness "the wisdom of cold weapon era". Most of the early object detection algorithms were built based on handcrafted features. Their performance easily stagnates by constructing complex ensembles which combine multiple low-level image features with high-level context from object detectors and scene classifiers.

3.1.1 Viola Jones Detectors

Developed by Paul Viola and Michael Jones back in 2001 [9], the Viola-Jones Object Detection Framework can quickly and accurately detect objects in images and works particularly well with the human face. The algorithm combines the concepts of Haar-like Features, Integral

Images, the AdaBoost Algorithm, and the Cascade Classifier in order to create a system for object detection that is fast and accurate.

The VJ detector follows a most straight forward way of detection, i.e, sliding windows: to go through all possible locations and scales in an image to see if any window might contains a human face. Althought it seems to be a very simple process, the calculation behind it was far beyond the computer's power of its time. The VJ detector has dramatically improved its detection speed by incorporating three important techniques: "integral image", "feature selection", and "detection cascades". While the VJ detector was tens or even hundreds of times faster than any other algorithms in its time under comparable detection accuracy, but in the scope of this report, I will not get deeper in detail of how VJ detector work.

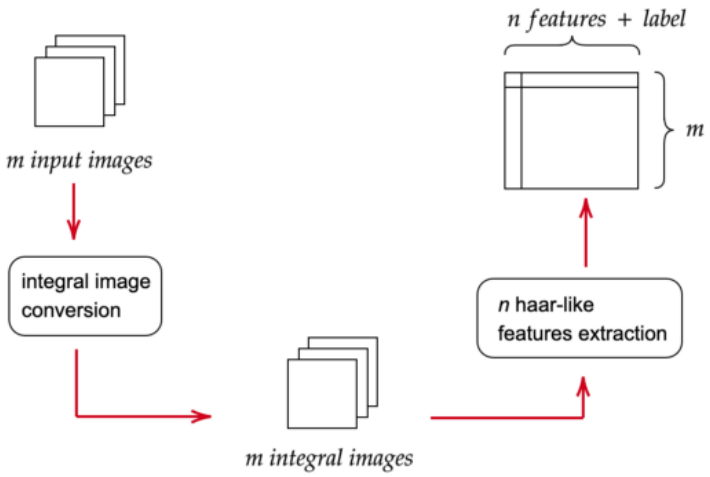

Figure 7: Integral image and how Haar exacting features

Source: *Socret Lee (towardsdatascience.com)*

3.1.2 Histogram of Oriented Gradients (HOG) Detector

Release by N. Dalal and B. Triggs in 2005 [11], the HOG methods, Histogram of Oriented Gradients, were considered as an important improvement of the scale-invariant feature transform and shape contexts of its time. The core of HOG algorithm is image gradient vector. It's a metric for every individual pixel, containing the pixel color changes in both x-axis and y-axis which represent the dirrection of colors changing from one extreme to the other.

The HOG descriptor is designed to compute the gradient vector of every pixel on a dense grid of uniformly spaced cells. Then it searchs for all possible regions that counld contains

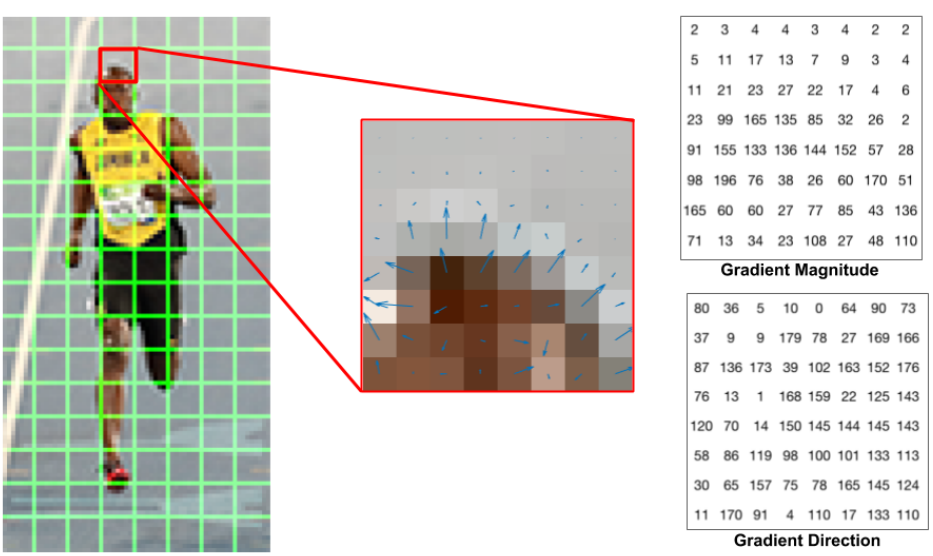

Figure 8: Visuallization of image gradient vector

Source: *learnopencv.com*

object in images and use overlapping local contrast normalization (on block) for improving accuracy. To able to detect objects of different size, the HOG detector rescales the input image for multiple times while keeping the size of detecting window unchanged.

3.1.3 Deformable Part-based Model (DPM)

DPM was originally proposed by P.Felzenswalb in 2008 [12] as an extension of the HOG detector. The PASCAL Visual Object Classes Challenge 2007, 2008 and 2009 pronounce DPM as the winners, therefore, it can be called as the peak of the traditional object detection methods.

Followings the detection philosophy of "divide and conquer", where the training can be simply considered as the learning of a proper way of decomposing an object. and the inference can be considered as an ensemble of detections on different object parts. For example, when detecting a "bicycle" can be considered of detecting wheels, body. This was named "star-model" published by P. Felzenszwalb [12].

A typical DPM detector consists of a root-filter and a number of part-filters. Instead of manually specifying the configurations of the part filters. Instead of manually specifying the configurations of the part filters (e.g., size, location), a weakly supervised learning method is developed in DPM where all configurations of part filters can be learned automatically.

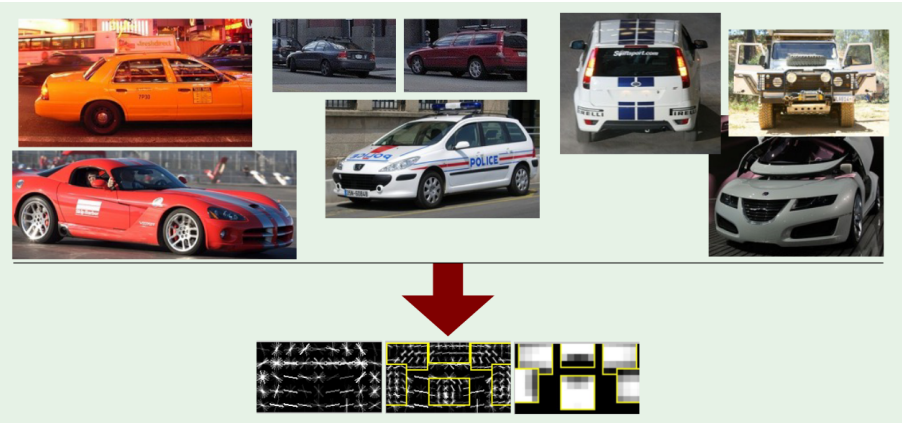

Figure 9: Root-filter and part-filters of a car in DPM algorithm

Source: *CS231b Stanford*

3.2 Object Detection based Deep Learning

The Deep learning evolution have started when AlexNet[23] took the first place in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. At that time, the convolutional operations was not well known, although, the concept of it's first publish in 1998 by LeCun, one of the "god father" of AI. AlexNet was a mystery or dark magic back there, while the first place of ImageNet 2011 achieved 25.8% in top 5 errors, AlexNet has make only 16.4% of error, which is approximately a half of last highest score. People said that because of the appearance of AlexNet, the world saw the rebirth of convolutional neural networks.

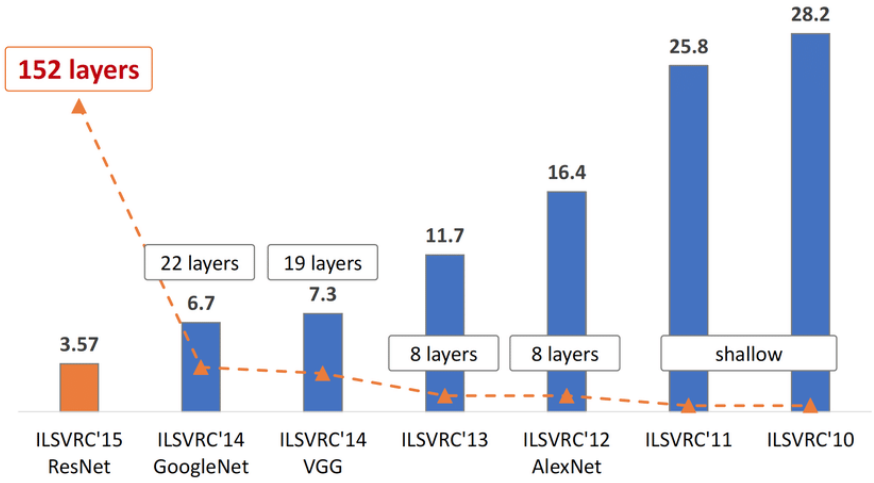

Figure 10: An overview of ImageNet in 5 years

Source: *Kien-Nguyen-23 at researchgate.net*

Only 4 years after Krizhevsky, Sutskever, Hinto publish the archtecture of AlexNet[23], the first CNN model, Kaiming He represent ResNet[24], a new approach to Deep Learning, achieved only 3.6% in top 5 error which is higher than human score. The Computer Scientists in general and scientists interested in Object detection started wonder how CNN or Deep Learning can be apply to CV tasks. While the performance of hand-crafted features became saturated, deep convolutional network is able to rebust and high-level feature representations of an image. Since then, object detection started to evolve at an unprecedented speed, many efficient models were presented for solving Object Detection such as: R-CNN[15], YOLO[19], SSD[20], etc.

3.2.1 CNN based Two-stage Detectors

In deep learning era, object detection can be grouped into two genres: "two stage detection" and "one-stage detection", where the former frames the detection as a "coarse-to-fine" process while the later frames it as to "complete in one step". The group of "two stage detection" has R-CNN, Fast R-CNN, Faster R-CNN as distinguished characteristics, however, in this report, I will discuss only about the basic concept of R-CNN family through first R-CNN model.

RCNN

The idea behind RCNN is simple: It starts with the extraction of a set of object proposal (object candidate boxes) by selective search. Then each proposal is rescaled to a fixed size image and fed into a CNN model trained on a pretrained model such as: AlexNet, VGG16, etc, to extract features.

Figure 11: A Graph Based Image Segmentation

Finally, linear SVM classifiers are applied to predict the presence of an object within each region and to recognize object categories. As one of the first model based on Deep Learning,

RCNN maked a signicant performance boost on VOC07 with a large improvement of mAP (mean Average Precision) from 33.7% by DPM [12] to 58.5%.

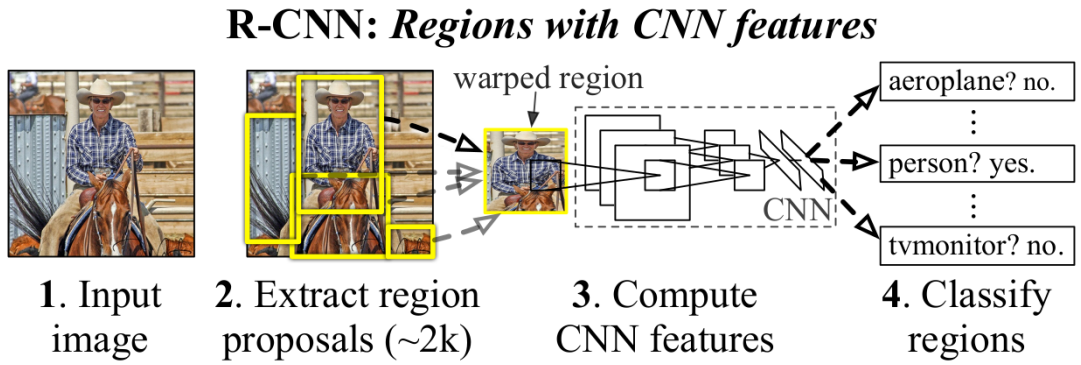

Figure 12: An overview of R-CNN system

Source: *Rich feature hierarchies for accurate object detection and semantic segmentation [15]*

3.2.2 CNN based One-stage Detectors

You Only Look Once (YOLO)

YOLO was proposed by R. Joseph *et al.* in 2015 [19] as the first one-stage detector. The biggest advantage of YOLO than R-CNN that is extremely fast. A fast version of YOLO runs at 155fps with VOC07 mAP = 52.7%. YOLO is based on the idea of segmenting an image into smaller images. This network divides the image into regions and predicts bounding boxes and probabilities for each region (figure 13). These bounding boxes are weighted by the predicted probabilities. Later, R. Joseph has made a series of improvements on basis of YOLO and has proposed its v2 and v3 editions [25, 26], which improve the detection accuracy while keeps a very high detection speed.

Single Shot MultiBox Detector (SSD)

As the second one-stage detector, SSD [20] was proposed by W. Liu *et al.* in 2015. The model discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes.

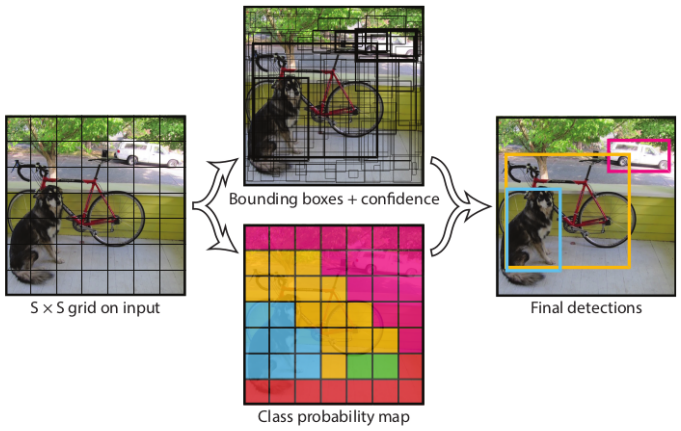

Figure 13: YOLO model

Source: *You Only Look Once [19]*

The main contribution of SSD is the introduction of the multi-reference and multi-resolution detection techniques, which significantly improves the detection accuracy of a one-stage detector, especially for small objects which is the original YOLO weakness. SSD jas asdvamtages om terms of both detection speed and accuracy (VOC07 mAP=76.8%, VOC12 mAP=74.9%, COCO mAP@.5=46.5%, mAP@.5:.95=26.8%, a fast version runs at 59fps).

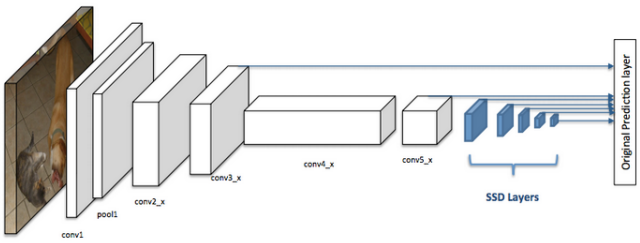

Figure 14: Architecture of a convolutional neural network with a SSD detector

Source: *SSD: Single shot multibox detector [20]*

REFERENCES

[1] Tracita. *Artificial Intelligence statistics*. URL: https://www.tractica.com/newsroom/press-releases/computer-vision-hardware-and-software-market-to-reach-48-6-billion-by-2022/.

[2] Marvin Minsky Seymour Papert. *The Summer Vision Project*. Tech. rep. MIT Artificial Intelligence, 1966. URL: http://people.csail.mit.edu/brooks/idocs/AIM-100.pdf.

[3] AI Newsletter. "AI Expert Newsletter: W is for Winter". In: (9 November 2013).

[4] Kush Agrawal. *To study the phenomenon of the Moravec's Paradox*. 2010. arXiv: 1012.3148 [cs.AI].

[5] Wikipedia. *Deep Blue versus Garry Kasparov*. URL: https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov.

[6] M. A. Hearst et al. "Support vector machines". In: *IEEE Intelligent Systems and their Applications* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428.

[7] Larry R Medsker and LC Jain. "Recurrent neural networks". In: *Design and Applications* 5 (2001).

[8] Zhengxia Zou et al. "Object detection in 20 years: A survey". In: *arXiv preprint arXiv:1905.05055* (2019).

[9] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. IEEE. 2001, pp. I–I.

[10] Paul Viola and Michael J Jones. "Robust real-time face detection". In: *International journal of computer vision* 57.2 (2004), pp. 137–154.

[11] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, pp. 886–893.

[12] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model". In: *2008 IEEE conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[13] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. "Cascade object detection with deformable part models". In: *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 2241–2248.

[14] Pedro F Felzenszwalb et al. “Object detection with discriminatively trained part-based models”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.9 (2009), pp. 1627–1645.

[15] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[16] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[17] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. “A-fast-rcnn: Hard positive generation via adversary for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2606–2615.

[18] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv preprint arXiv:1506.01497* (2015).

[19] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[20] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[21] Tsung-Yi Lin et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[22] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[24] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[25] Joseph Redmon and Ali Farhadi. *YOLO9000: Better, Faster, Stronger*. 2016. arXiv: 1612.08242 [cs.CV].

[26] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV].