How Do Intelligent Systems Recognize and Track Objects Using Neural Networks?

COMP 3190

ALBORZ KHAKBAZAN

Abstract

Computer vision has been a difficult subject of research for many years, but over the past decade, it is steadily becoming more and more powerful as new approaches are developed, which bring us closer to truly creating vision that is both highly efficient and accurate. However, in order to achieve this, we must further develop computer vision to mimic that of humans, by relaying sensory inputs to the brain. By using artificial intelligence and neural networks in computer vision, we can see some incredible future applications of this technology to benefit society. This research paper is intended to delve deeper into computer vision and explore why artificial intelligence and neural networking is crucial for object detection, how it is utilized, and what the future for it looks like in our ever-changing world.

Section 1: Introduction

1.1 Basics of Computer Vision

There are a variety of different tasks in computer vision, ranging in levels of complexity. At its most basic, computer vision exists in the form of image classification, where the system receives an image and simply classifies the type of objects that exists within the scene, and object localization, where the system locates and places a bounding box around an object in a scene. [Zhang et al. 2013] While these have their uses, computer vision can be taken a step further in the form of object detection, which allows us to spatially locate and classify multiple different objects in a scene by placing labelled bounding boxes around them and updating them frame-by-frame. We can go even further than that using object segmentation, which locates and classifies objects by the pixel, aiming for complete accuracy in vision.

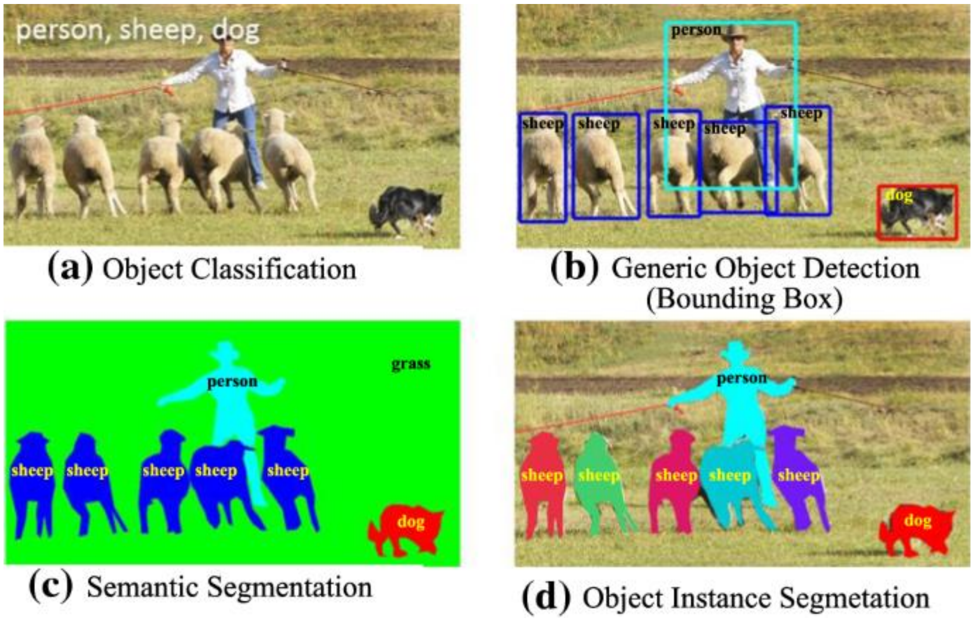

Figure 1

In an ideal world, computer vision is just that – completely accurate, high-speed vision that surpasses that of our own. However, we are still far and away from perfecting computer vision – at its lowest level, it is highly efficient but highly inaccurate, and at its greatest, it is incredibly accurate while also incredibly slow. Object detection is a form of computer vision that manages to balance accuracy with efficiency in a way segmentation has yet to do, and so it is the main form of computer vision that this paper will delve into. This form of vision has been advancing rapidly in the past few years, becoming exponentially faster as computer scientists find new ways to approach object detection, but it could not have been done without artificial intelligence.

1.2 Why is AI Necessary for Vision?

As far as vision goes, it would be wonderful for our systems if every object looked exactly the same 100% of the time. With one reference frame or image, you could classify an object no matter where it appeared or what the scene as a whole looked like. Unfortunately, we live in a real world, with a plethora of factors that can change the appearance of a scene – such as the lighting, scale, clutter, blurriness, object poses, and many more. [Liu et al. 2020] These factors can be described as imaging conditions, but images of the same class can also vary by intrinsic factors. Objects are unique and distinct - people can have different facial features, cars can have different makes or models, and even streetlights can have different designs.

Take something as simple as a dog and try to explain its appearance – well, you could describe them as a four-legged furry creature, with floppy ears, a wide mouth, a snout nose, and a long, slender tail. For humans, this is an adequate description – but our computer systems lack the familiarity we have with the world that makes this description so accurate. What about variations in different dog breeds, such as pugs? Their noses look quite different. Shih Tzus? Well, their tails aren't exactly long and slender. By our description, they should not be dogs… yet they are, and we know they are. There are features here that we recognize naturally and unconsciously, features that are subtle and practically impossible to describe in words such as bone structure or skin texture. We have *learned* that these features are what make a dog look like a dog, using a neat little information center called the brain which ties all of our knowledge together with our vision and help us make sense of what we see.

This is exactly why artificial intelligence is so important for accurate computer vision. Although computer systems are not necessarily trying to replicate human vision, by using deep learning and neural networks, computers too can have a “brain” of their own so that they may learn features and efficiently detect objects.

Section 2: Deep Learning in Object Detection

2.1 CNNs

Convolutional Neural Networks, commonly referred to as CNNs, can be described most primitively as an artificial brain, the centerpiece for deep learning in our computer systems. They are hierarchical and complex, comprising of a multitude of different layers depending on how robust the network aims to be, where each layer in the CNN is comprised of feature maps or pattern stores for object classes. [Lecun et al. 2015; Liu et al. 2020] Deep CNNs, or DCNNs, have a very large amount of layers, and excel in their ability learn object features with minimal exposure, as well as process complex functions that a smaller CNN would not be able to. However, DCNNs are limited in that they also require training data and computing power that is just as robust.

2.2 Edge Detection

In order to detect patterns, CNNs build feature maps on each convolutional layer using edge detection. There are different approaches to this, but generally feature maps are built by constructing a “filter” or “kernel” matrix to convolve over an input frame’s matrix representation. [Canny 1986] The output becomes a new image matrix representation, where lower values in the matrix correspond to edges in the image. In fig. 2, we see various different edge detection outputs in their image representation, which use 2 different filter matrixes to detect horizontal edges and vertical edges before combining the two. [Jain et al. 1995]

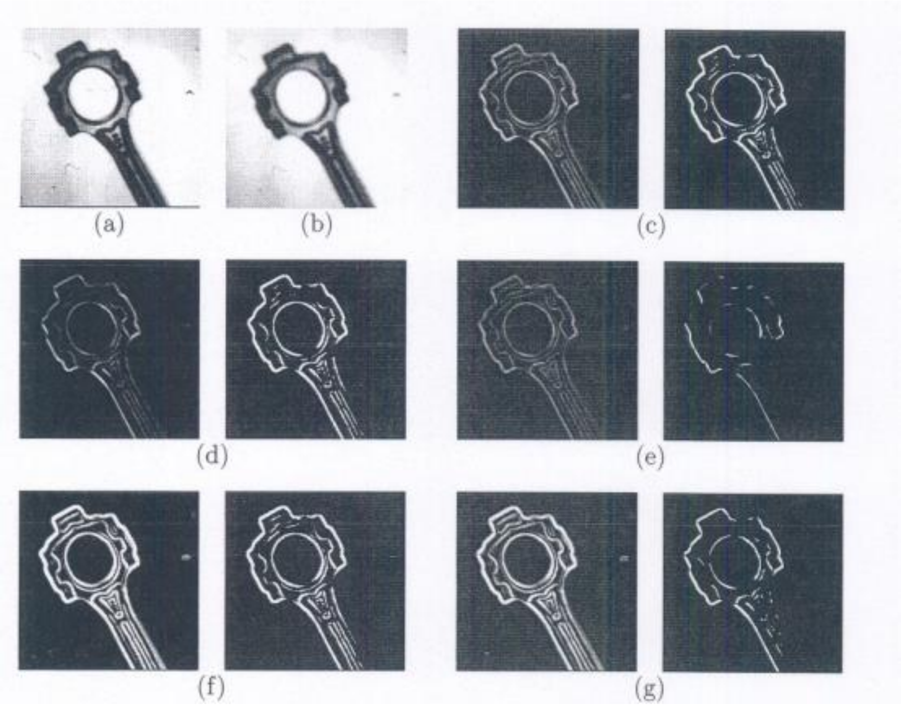

Figure 2

Feature maps are built from a wide array of different patterns in order to cover both intrinsic class variation and the effect of imaging conditions. With potentially millions of different patterns to refer to, the system can accurately isolate then exact features that make up a class and use them to detect objects in the real world.

2.3 Training

So, we have CNNs to detect patterns using edge detection, and to store this information for the computer to refer to – but where exactly do these patterns come from? If the system doesn't know what an object looks like, how will it detect it in the real world? There are currently four major datasets which exist both for computer vision training, as well as performance evaluation when comparing different algorithms. These datasets consist of collections of annotated images or frames which each have their own respective challenges and uses in the world of object detection.

The first, PASCAL VOC, comprises of images in 20 different object categories with bounding box labels and was one of the benchmark datasets that set the precedent for future computer vision evaluation in 2005. [Everingham et al. 2015] While it has since been out shadowed by larger, more modern datasets, it still has its uses in simple object detection with everyday, common items.

The second, ILSVRC, or ImageNet, is much larger with 1000 different object categories and 1.2 million images, using bounding box labels like PASCAL VOC. [Deng et al. 2009] However, one issue with ImageNet is its usage of images that are fairly straightforward and well-centered – as mentioned earlier, the appearance of objects in the real world is affected heavily by the surrounding environment, leading researchers to develop MS COCO.

MS COCO provides bounded or segmented images of objects in their natural environments, usually amid heavy clutter or in more typical real-world lighting. [Lin et al. 2014] More intricate scenes, as well as more intricate object labels provide the system with better feedback to learn have led to MS COCO being the standard for object detection today.

Lastly, we have OICOD, or OpenImages. Like MS COCO and ImageNet, OpenImages uses bounding box labels and works at a large-scale, but OpenImages shines in that it contains significantly more annotations and images than any other dataset available to the public. [Kuznetsova et al 2018]

While MS COCO is the most relevant and commonly used dataset today, all 4 have their place in object detection and have their role leading to the developments that have come since.

Section 3: Approaching Object Detection

3.1 Region-Based Frameworks

An early approach that played a crucial role in object detection's progress today, R-CNN frameworks are a method for object detection where the system computes and returns regions in an image that likely contain objects using a selective search, then crops and warps these regions to size before sending them to a trained CNN to confirm that the region contains an object. The image is then classified, and finally a bounding box is placed over the object. [Girshick et al 2014]

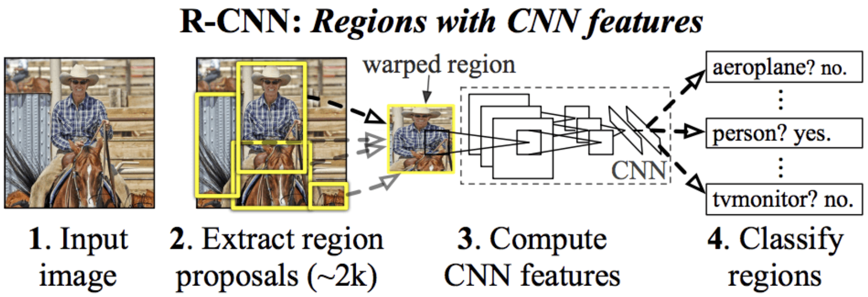

Figure 3

At its time, R-CNN was groundbreaking – but since R-CNN is a multistage framework with functions in different stages, it was slow and costly in storage, leading to challenges in training the system as well as in large-scale object detection. It was improved upon a number of times throughout the years – namely, in the form of “Fast R-CNN” and “Faster R-CNN”. Fast R-CNN aimed to improve detection speed by using “Region of Interest Pooling” in order to extract features from each region proposal. This allowed Fast R-CNN to simultaneously train its image classification as well its bounding box regression and improve efficiency by 3 times in training and up to 10 times in regular use. Faster R-CNN improved upon this efficiency even further by employing its CNN to replace the selective search that returned object region proposals, with a “Region Proposal Network” that shared convolutions with the CNN. [Ren et al. 2016]

3.2 You Only Look Once

As effective and simplified as R-CNN has become, it is still at its core a multistage framework. Regions must first be found by the RPN before placing a bounding box and classifier on the object. YOLO, or “You Only Look Once”, is a framework that aims to complete object detection within a single stage. While R-CNN finds local features within each region, revisiting the same image hundreds of times to check all possible regions, YOLO instead opts to divide the input into an SxS grid and make class proposals as well as bounding box proposals simultaneously. [Redmon et al. 2015] If a grid cell contains the center of a bounding box, it is proposed along with its confidence that it contains an object, along with the confidence levels for the possible object class for that cell.

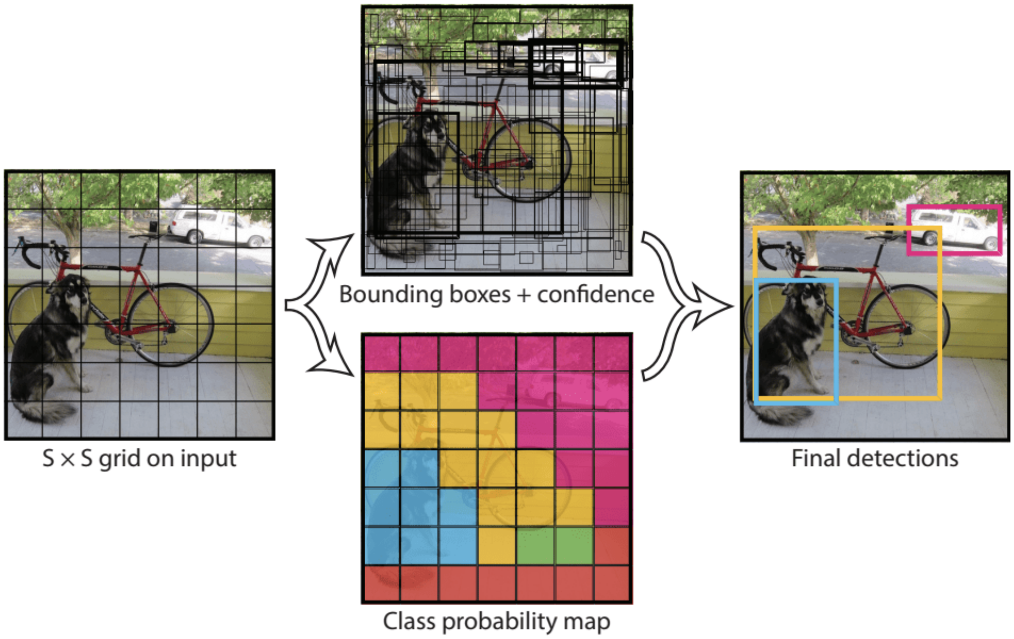

Figure 4

Although this method was initially less accurate than the R-CNN approach when it was proposed in 2015, it has since been improved with v2 in 2016 and v3 in 2018 which have made it the most promising framework for object detection today. It is incredibly efficient - running reasonably accurate object detection at up to 155 frames per second. YOLO is far from perfect still – with its efficiency, it has potential to pave the way for all sorts of advanced vision technology using live video feedback, but its accuracy still limits it from being used confidently there. With a few more years of research and development, YOLO can hopefully push us forward to the future for computer vision.

Section 4: Discussion

4.1 Challenges in Computer Vision

As promising as YOLO currently is, there are still limitations in computer vision in the modern day. In an ideal world, computer vision is both highly accurate and highly efficient, and perhaps someday in the future we can develop approaches to object segmentation that are as important as YOLO is for detection. However, it is impossible to maximize vision efficiency without making at least some sort of trade off in accuracy – we simply cannot maximize speed and minimize memory use while distinctly recognizing and tracking large numbers of objects in the real world. [Zhang et al. 2013]

Even just developing a system that is accurate is inherently difficult – we simply cannot have vision that can deal with intra-class variation to its fullest potential while also dealing with inter-class variation at the same level. Having a system that is incredibly robust in the former means being able to handle any sort of imaging condition, image distortion, or intrinsic factor for an object class. When we can potentially have thousands of different object classes, being able to definitively handle every single one of these classes simply requires too much computing power. On top of that, the accuracy of our systems is limited by the complexity of the CNNs we use – which are in turn limited by the availability of viable training data and our computing power. As always, we can make strides in computer technology that will help us come closer to this ideal object vision, but there are challenges we face and trade-offs we must make before we get anywhere near.

4.2 The Future of Computer Vision

Computer vision has made a great deal of progress in the past few years. Technology that seemed light-years away is now becoming ever more tangible – like self-driving vehicles, automatic image captioning, and biometric phone locks. However, the potential for computer vision is so grand, and not just limited to the consumer – there are also a great deal of possible applications that would push our society forward and benefit everybody, even the less fortunate. In the medical field, researchers have developed an AI that identifies cancer tumors from CT scan images, yielding better results than human radiologists. [Liu et. al 2020] For security, surveillance cameras are becoming more and more practical without having a human monitoring them, and the progress we make in computer vision that deals with large amounts of clutter could detect shooters or terrorist attacks early, automating a response before people are hurt. There is even potential for computer vision to substitute vision for the visually impaired as we continue to progress this technology, using live image captioning to describe a person's surroundings as they go about their day.

Coupled with other artificial intelligence technologies, the potential for computer vision is vast, and still largely unexplored. As we develop better approaches to object detection or segmentation, more powerful computers, and greater collections of training data for DCNNs, we can start to see bits and pieces of what the potential for advanced computer vision looks like.

Bibliography

- Diamant E. (2005) Does a Plane Imitate a Bird? Does Computer Vision Have to Follow Biological Paradigms?. In: De Gregorio M., Di Maio V., Frucci M., Musio C. (eds) Brain, Vision, and Artificial Intelligence. BVAI 2005. Lecture Notes in Computer Science, vol 3704. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11565123_11
- Liu, L., Ouyang, W., Wang, X. *et al.* Deep Learning for Generic Object Detection: A Survey. *Int J Comput Vis* 128, 261–318 (2020). https://doi.org/10.1007/s11263-019-01247-4
- Recognition by Top-Down and Bottom-Up Processing in Cortex: The Control of Selective Attention
  Dan Graboi and John Lisman, Journal of Neurophysiology 2003 90:2, 798-810
- llman, S. (2000). *High-level Vision: Object Recognition and Visual Cognition*. MIT Press.
- Bennamoun, M., & Mamic, G. (2012). *Object Recognition: Fundamentals and Case Studies*. Springer London.
- Canny, J., A Computational Approach To Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986
- Jain, R. et al. (1995) Machine Vision, McGraw-Hill, Inc.
- Zhang, X., Yang, Y., Han, Z., Wang, H., & Gao, C. (2013). Object class detection: A survey. *ACM Computing Surveys*, *46*(1), 10:1–10:53.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *IJCV*, *115*(3), 211–252.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.
- Everingham, M., Eslami, S., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *IJCV*, *111*(1), 98–136.
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). ImageNet: A large scale hierarchical image database. In *CVPR* (pp. 248–255).
- Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, L. (2014). Microsoft COCO: Common objects in context. In *ECCV* (pp. 740–755).
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., PontTuset, J., et al. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982.

- Ren S. et al. (2016).“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” (2016)

- Girshick, R. et al. (2014) “Rich feature hierarchies for accurate object detection and semantic segmentation.”

- Joseph, R. et al. (2015) You Only Look Once: Unified, Real-Time Object Detection,