1 Introduction to the World of Vision

Supplementary content at http://bit.ly/2TqTDt5

Understanding how the brain works constitutes the greatest scientific challenge of our times, arguably the greatest challenge of all times. We have sent spaceships to peek outside of our solar system, and we study galaxies far away to build theories about the origin of the universe. We have built powerful accelerators to scrutinize the secrets of subatomic particles. We have uncovered the secrets to heredity hidden in the billions of base pairs in DNA. But we still have to figure out how the three pounds of brain tissue inside our skulls work to enable us to do physics, biology, music, literature, and politics.

The conversations and maneuvers of about a hundred billion neurons in our brains are responsible for our ability to interpret sensory information, to navigate, to communicate, to feel and to love, to make decisions and plans for the future, to learn. Understanding how neural circuits give rise to cognitive functions will transform our lives: it will help us alleviate the ubiquitous mental health conditions that afflict hundreds of millions, it will lead to building true artificial intelligence machines that are as smart as or smarter than we are, and it will open the doors to finally understanding who we are.

As a paradigmatic example of brain function, we will focus on one of the most exquisite pieces of neural machinery ever evolved: the visual system. In a small fraction of a second, we can get a glimpse of an image and capture a substantial amount of information. For example, we can take a look at Figure 1.1 and answer an infinite series of questions about it: *who is there, what is there, where is this place, what is the weather like, how many people are there, and what are they doing?* We can even make educated guesses about a potential narrative, including describing *the relationship between people in the picture,* w*hat happened before*, or *what will happen next*. At the heart of these questions is our capacity for visual cognition and intelligent inference based on visual inputs.

Our remarkable ability to interpret complex spatiotemporal input sequences, which we can loosely ascribe to part of "common sense," does not require us to sit down and solve complex differential equations. In fact, a four-year-old can answer most of the questions outlined before quite accurately, younger kids can answer most of them, and many non-human animal species can also be trained to correctly describe many aspects of a visual scene. Furthermore, it takes only a few hundred milliseconds to deduce such profound information from an image. Even though we have computers that excel at

Figure 1.1 We can visually interpret an image at a glance. Who is there? What is there? Where is it? What are they doing? What will happen next? These are just examples among the infinite number of questions that we can answer after a few hundred milliseconds of exposure to a novel image.

tasks such as solving complex differential equations, computers still fall short of human performance at answering common-sense questions about an image.

1.1 Evolution of the Visual System

Vision is essential for most everyday tasks, including navigation, reading, and socialization. Reading this text involves identifying shape patterns. Walking home involves detecting pedestrians, cars, and routes. Vision is critical to recognize our friends and decipher their emotions. It is, therefore, not much of a strain to conceive that the expansion of the visual cortex has played a significant role in the evolution of mammals in general and primates in particular. It is likely that the evolution of enhanced algorithms for recognizing patterns based on visual inputs yielded an increase in adaptive value through improvements in navigation, discrimination of friend versus foe, differentiation between food and poison, and through the savoir faire of deciphering social interactions. In contrast to tactile and gustatory inputs and, to some extent, even auditory inputs, visual signals bring knowledge from vast and faraway areas. While olfactory signals can also diffuse through long distances, the speed of propagation and information content is lower than that of photons.

The ability of biological organisms to capture light is ancient. For example, many bacteria use light to perform photosynthesis, a precursor to a similar process that

captures energy in plants. What is particularly astounding about vision is the possibility of using light to capture *information* about the world. The selective advantage conveyed by visual processing is so impactful that it has led the zoologist Andrew Parker to propose the so-called light switch theory to explain the rapid expansion in number and diversity of life on Earth.

About five hundred million years ago, during the early Cambrian period, there was an extraordinary outburst in the number of different species. It is also at around the same time where fossil evidence suggests the emergence of the first species with eyes, the trilobites (Figure 1.2). Trilobites are extinct arthropods (distant relatives of insects and spiders) that conquered the world and expanded throughout approximately three hundred million years. The light switch theory posits that the emergence of eyes and the explosion in animal diversity is not a mere coincidence. Several investigators have argued that eyes emerged right before the Cambrian explosion. Eyes enabled some lucky early trilobite, or its great grandfather, to capture information from farther away, detecting the presence of prey or predator, endowing it with a selective advantage over other creatures without eyes, who had to rely on slower and coarser information for survival. Using these new toys, the eyes, an evolutionary arms race commenced between prey and predators to make inferences about the world around them and to hide from those scrutinizing and powerful new sensors. All of a sudden, body shapes, textures, and colors became fascinating, powerful, and dangerous. It seems likely that

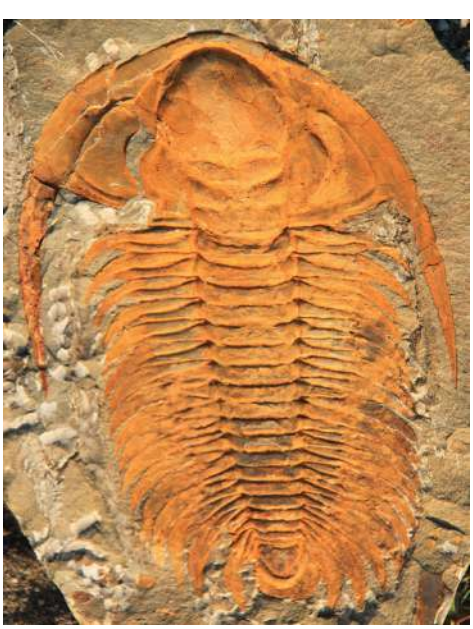

Figure 1.2 Fossil record of a trilobite, circa 500 million years ago. Trilobites such as the one shown in this picture had compound eyes, probably not too different from those found in modern invertebrate species like flies. Trilobites proliferated and diversified throughout the world for about 300 million years. By Dwergenpaartje, CC BY-SA 3.0

body shapes and colors began to change to avoid detection through the initial versions of camouflage – in turn, leading to keener and better eyes to be more sensitive to motion and to subtle changes through the ability to better discriminate shapes. Let there be light, and let light be used to convey information.

1.2 The Future of Vision

Fast-forward several hundred million years, the fundamental role of vision in human evolution is hard to underestimate. Well before the advent of language as it is known today, vision played a critical role in communication, interpreting emotions and intentions, and facilitating social interactions. The ability to visually identify patterns in the position of the moon, the sun, and the stars led to understanding and predicting seasonal changes, which eventually gave rise to agriculture, transforming nomadic societies into sedentary ones, begetting the precursors of future towns. Art, symbols, and eventually the development of written language also fundamentally relied on visual pattern-recognition capabilities.

The evolution of the visual system is only poorly understood and remains an interesting topic for further investigation. The future of the visual system will be equally fascinating. While speculating about the biological changes in vision in animals over evolutionary time scales is rather challenging, it is easier to imagine what might be accomplished in the near future over shorter time scales, via machines with suitable cameras and computational algorithms for image processing. We will come back to the future of vision in Chapter 9; as a teaser, let us briefly consider machines that can achieve, and perhaps surpass, human-level capabilities in visual tasks. Such machines may combine high-speed and high-resolution video sensors that convey information to computers implementing sophisticated simulations that approximate the functions of the visual brain in real time.

Machines may soon excel in face-recognition tasks to a level where an ATM will greet you by your name without the need for a password, where you may not need a key to enter your home or car, where your face may become your credit card and your passport. Self-driving vehicles propelled by machine vision algorithms have escaped the science fiction pages and entered our streets. Computers may also be able to analyze images intelligently to search the web by image or video content (as opposed to keywords and text descriptors). Doctors may rely more and more on artificial vision systems to analyze X-rays, MRIs, and other images, to a point where image-based diagnosis becomes the domain of computer science entirely. Future generations may be intrigued by the notion that we once let fallible humans make diagnostic decisions. The classification of distant galaxies, or the discovery of different plant and animal species, might be led by machine vision systems rather than astronomers or biologists.

Adventuring further into the domain of science fiction, one could conceive of brain–machine interfaces that might be implanted in the human brain to augment visual capabilities for people with visual impairment or blind people. While we are at it, why not also use such interfaces to augment visual function in normally sighted people

to endow humans with the capability to see in 360 degrees, to detect infrared or ultraviolet wavelengths, to see through opaque objects such as walls, or even directly witness remote events?

When debates arose about the possibility that computers could one day play competitive chess against humans, most people were skeptical. Simple computers today can beat even sophisticated chess aficionados, and good computers can beat world champions. Recently, computers have also excelled in the ancient and complex game of Go. Despite the obvious fact that most people can recognize objects much better than they can play chess or Go, visual cognition is actually more challenging than these games from a computational perspective. However, we may not be too far from building accurate computational approximations to visual systems, where we will be able to trust computers' eyes as much as, or even more than, our own eyes. Instead of "seeing is believing," the future moto may become "computing is believing."

1.3 Why Is Vision Difficult?

The notion that seeing is computationally more complicated than playing Go may be counterintuitive. After all, a two-year-old child can open her eyes and rapidly recognize and interpret her environment to navigate the room and grab her favorite teddy bear, which may be half-covered behind other toys. She does not know how to play Go. She certainly has not gone through the millions of hours of training via reinforcement learning that neural network machines had to go through to play Go. She has had approximately ten thousand hours of visual experience. These ten thousand hours are mostly unsupervised; there were adults nearby most of the time, but, by and large, those adults were not providing continuous information about object labels or continuous reward and punishment signals (there certainly were labels and rewards, but they probably constituted a small fraction of her visual learning).

Why is it so difficult for computers to perform pattern-recognition tasks that appear to be so simple to us? The primate visual system excels at recognizing patterns even when those patterns change radically from one instantiation to another. Consider the simple line schematics in Figure 1.3. It is straightforward to recognize those handwritten symbols even though, at the pixel level, they show considerable variation within each row. These drawings only have a few traces. The problem is far more complicated with real scenes and objects. Imagine the myriad of possible variations of pictures taken at Piazza San Marco in Venice (Figure 1.1) and how the visual system can interpret them with ease. Any object can cast an *infinite* number of projections onto the eyes. These variations include changes in scale, position, viewpoint, and illumination, among other transformations. In a seemingly effortless fashion, our visual systems can map all of those images onto a particular object.

Identifying specific objects is but one of the important functions that the visual system must solve. The visual system can estimate distances to objects, predict where objects are heading, infer the identity of objects that are heavily occluded or camouflaged, determine which objects are in front of which other objects, and make educated

Figure 1.3 Any object can cast an infinite number of projections onto the eyes. Even though we can easily recognize these patterns, there is considerable variability among different renderings of each shape at the pixel level.

guesses as to the composition and weight of objects. The visual system can even infer intentions in the case of living agents. In all of these tasks, vision is an ill-posed problem, in the sense that multiple possible solutions are consistent with a given pattern of inputs onto the eyes.

1.4 Four Key Features of Visual Recognition

In order to explain how the visual system tackles the identification of patterns, we need to account for four key features of visual recognition: *selectivity, tolerance, speed*, and *capacity.*

Selectivity involves the ability to discriminate among shapes that are very similar at the pixel level. Examples of the exquisite selectivity of the visual system include face identification and reading. In both cases, the visual system can distinguish between inputs that are very close if we compare them side by side at the pixel level. A trivial and useless way of implementing *Selectivity* in a computational algorithm is to memorize all the pixels in the image (Figure 1.4A). Upon encountering the same pixels, the computer would be able to "recognize" the image. The computer would be very selective because it would not respond to any other possible image. The problem with this implementation is that it lacks *tolerance.*

Tolerance refers to the ability to recognize an object despite multiple transformations of the object's image. For example, we can recognize objects even if they are presented in a different position, scale, viewpoint, contrast, illumination, or color. We can even recognize objects where the image undergoes nonrigid transformations, such as the changes a face goes through upon smiling. A straightforward but useless way of implementing tolerance is to build a model that will output a flat response no matter the input. While the model would show tolerance to image transformations, it would not

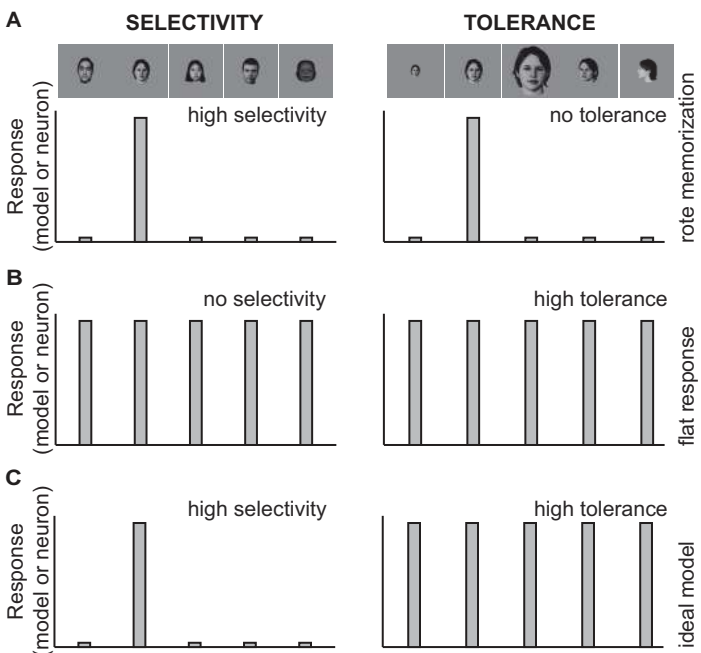

Figure 1.4 A naïve (and not very useful) approach to model visual recognition. Two simple models that are easy to implement, easy to understand, and not very useful. A rote memorization model (**A**) can have exquisite selectivity but does not generalize. In contrast, a flat response model (**B**) can generalize but lacks any selectivity. (**C**) An ideal model should combine selectivity and tolerance.

show any selectivity to different shapes (Figure 1.4B). Combining *selectivity* and *tolerance* (Figure 1.4C) is arguably the key challenge in developing computer vision algorithms for recognition tasks. To consider a real-world example, a self-driving car needs to selectively distinguish pedestrians from many other types of objects, no matter how tall those pedestrians are, what they are wearing, what they are doing, or what they are holding.

Given the combinatorial explosion in the number of images that map onto the same "object," one could imagine that visual recognition requires many years of learning at school. Of course, this is far from the case. Well before a first grader starts to learn the basics of addition and subtraction (rather trivial problems for computers), she is already quite proficient at visual recognition, a task that she can accomplish in a glimpse. Objects can be readily recognized in a stream of other objects presented at a rate of 100 milliseconds per image. Subjects can make an eye movement to indicate the presence of an object in a two-alternative forced-choice task about 200 milliseconds after showing the visual stimulus. Furthermore, both scalp and invasive recordings from the human brain reveal signals that can discriminate among complex objects as early as ~150 milliseconds after stimulus onset. The *speed* of visual recognition constrains the number of computational steps that any theory of recognition can use to

account for recognition performance. To be sure, vision does not stop at 150 milliseconds. Many aspects of visual cognition emerge over hundreds of milliseconds, and recognition performance under challenging tasks improves with longer presentation times. However, a basic understanding of an image or the main objects within the image can be accomplished in ~150 milliseconds. We denote this regime as *rapid visual recognition.*

One way of making progress toward combining selectivity, tolerance, and speed has been to focus on object-specific or category-specific algorithms. An example of this approach would be the development of algorithms for detecting cars in natural scenes by taking advantage of the idiosyncrasies of cars and the scenes in which they typically appear. Another example would be face recognition. Some of these category- and context-specific heuristics are useful, and the brain may learn to take advantage of them. For example, if most of the image is blue, suggesting that the image background may represent the sky, then the prior probabilities for seeing a car would be low (cars typically do not fly), and the prior probabilities for seeing a bird would be high (birds are often seen against a blue sky). We will discuss the regularities in the visual world and the statistics of natural images in Chapter 2. Despite these correlations, in the more general scenario, the visual recognition machinery is capable of combining selectivity, tolerance, and speed for an enormous range of objects and images. For example, the Chinese language has more than three thousand characters. Estimations of the *capacity* of the human visual recognition system vary substantially across studies. Several studies cite numbers that are considerably more than ten thousand items.

In sum, a theory of visual recognition must be able to account for the high selectivity, tolerance, speed, and capacity of the visual system. Despite the apparent immediacy of seeing, combining these four key features is by no means a simple task.

1.5 The Travels and Adventures of a Photon

The challenge of solving the ill-posed problem of selecting among infinite possible interpretations of a scene in a transformation-tolerant manner within 150 milliseconds of processing seems daunting. How does the brain accomplish this feat? We start by providing a global overview of the transformations of visual information in the brain.

Light arrives at the retina after being reflected by objects in the environment. The patterns of light impinging on our eyes are far from random, and the natural image statistics of those patterns play an important role in the development and evolution of the visual system (Chapter 2). In the retina, light is transduced into an electrical signal by specialized photoreceptor cells. Information is processed in the retina through a cascade of computations before it passes on to a structure called the thalamus and, from there, on to the cortical sheet. The cortex directs the sequence of visual computation steps, converting photons into percepts. Several visual recognition models treat the retina as analogous to the pixel-by-pixel representation in a digital camera. A digital

camera is an oversimplified description of the computational power in the retina, yet it has permeated into the general jargon as introduced by manufacturers who boast of a "retina display" for monitors.

It is not unusual for commercially available monitors these days to display several million pixels. Commercially available digital cameras also boast tens of millions of pixels. The number of pixels in such devices approximates or even surpasses the number of primary sensors in some biological retinas; for example, the human retina contains ~6.4 million so-called cone sensors and ~110 million so-called rod sensors (we will discuss those sensors in Chapter 2). Despite these technological feats, electronic cameras still lag behind biological eyes in essential properties such as luminance adaptation, motion detection, focusing, energy efficiency, and speed.

The output of the retina is conveyed to multiple areas, including the superior colliculus, the suprachiasmatic nucleus, and the thalamus. The superficial layers of the superior colliculus can be thought of as an ancient visual brain. Indeed, for many species that do not have a cortex, the superior colliculus (referred to as optic tectum in these species) is where the main visual elaborations of the input take place. The suprachiasmatic nucleus plays a central role in regulating the circadian rhythm. Humans have an internal daily clock that runs slightly longer than the usual 24-hour day, and light inputs via the suprachiasmatic nucleus help modulate and adjust this cycle.

The main visual pathway carries information from the retina to a part of the thalamus called the lateral geniculate nucleus (LGN). The LGN projects to the primary visual cortex, located in the back of our brains. Without the primary visual cortex, humans are mostly blind, highlighting the critical importance of the pathway conveying information into the cortex for most visual functions. Investigators refer to the processing steps in the retina, LGN, and primary visual cortex as "early vision" (Chapter 5). The primary visual cortex is only the first stage in the processing of visual information in the cortex. Researchers have discovered tens of areas responsible for different aspects of vision (the actual number is still a matter of debate and depends on what is meant by "area"). An influential way of depicting these multiple areas and their interconnections is the diagram proposed by Felleman and Van Essen, shown in Figure 1.5. To the untrained eye, this diagram appears to depict a bewildering complexity, not unlike the circuit diagrams typically employed by electrical engineers. We will delve into this diagram in more detail in Chapters 5 and 6 and discuss the areas and connections that play a crucial role in visual cognition.

Despite the apparent complexity of the neural circuitry in Figure 1.5, this scheme is an oversimplification of the actual wiring diagram. First, each of the boxes in this diagram contains millions of neurons. There are many different types of neurons. The arrangement of neurons within each box can be described in terms of six main layers of the cortex (some of which have different sublayers) and the topographical arrangement of neurons within and across layers. Second, we are still far from characterizing *all* the connections in the visual system. One of the exciting advances of the last decade is the development of techniques to scrutinize the connectivity of neural circuitry at high resolution and in a high-throughput manner.

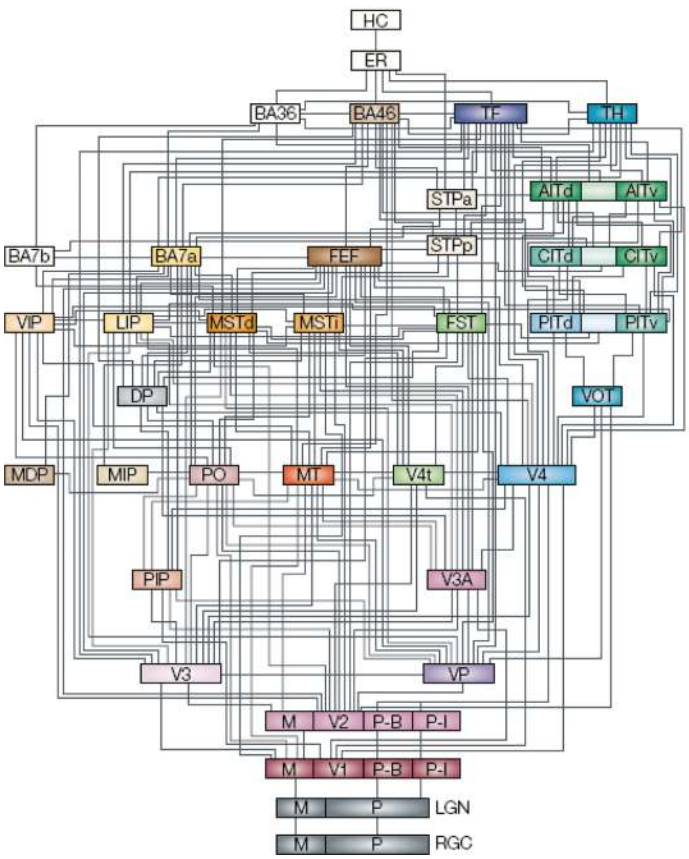

Figure 1.5 The adventures of a photon. Schematic diagram of the connectivity in the primate visual system. Adapted from Felleman and Van Essen 1991

For a small animal like a one-millimeter worm with the fancy name of *Caenorhabditis elegans*, we have known for a few decades the detailed connectivity pattern of each one of its 302 neurons, thanks to the work of Sydney Brenner (1927–2019). However, the cortex is an entirely different beast, with a neuronal density of tens of thousands of neurons per square millimeter. Heroic efforts in the burgeoning field of "connectomics" are now providing the first glimpses of which neurons are friends with which other neurons in the cortex. Major surprises in neuroanatomy will likely come from the use of novel tools that take advantage of the high specificity of molecular biology.

Finally, even if we did know the connectivity of every single neuron in the visual cortex, this knowledge would not immediately reveal the computational functions (but knowing the connectivity would still be immensely helpful). In contrast to electrical circuits where we understand each element and the overall function can be appreciated by careful inspection of the wiring diagram, many neurobiological factors make the map from structure to function a nontrivial one.