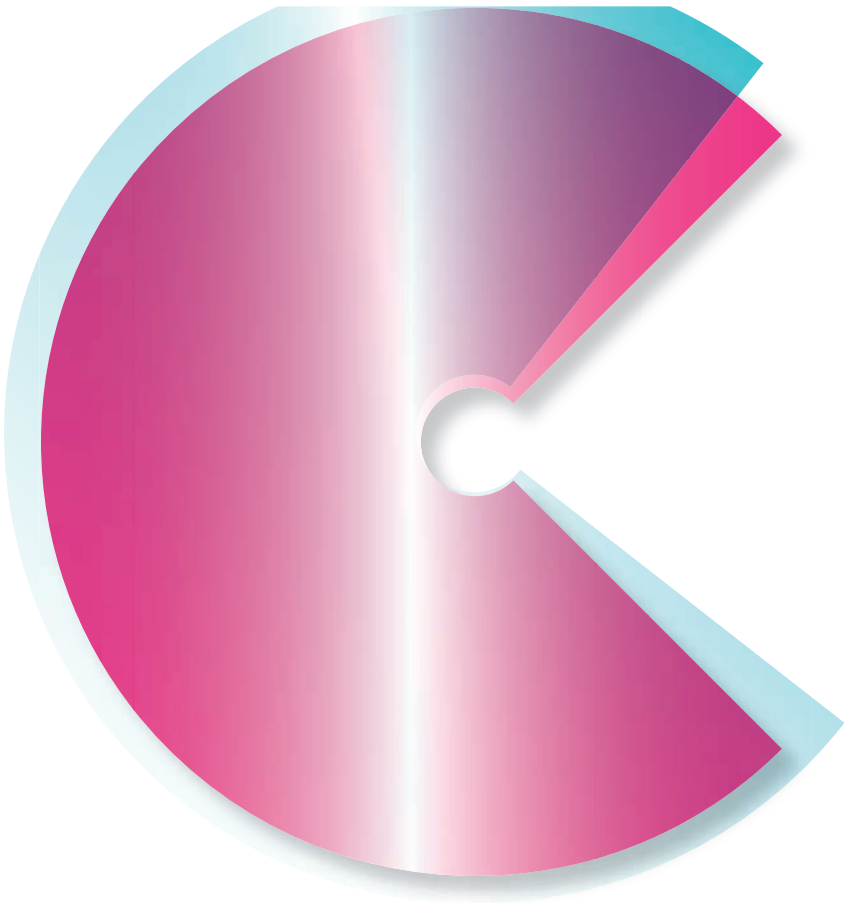
COMPUTER SCIENCE

MACHINES WHO LEARN

After decades of disappointment, artificial intelligence is finally catching up to its early promise, thanks to a powerful technique called deep learning

By Yoshua Bengio



COMPUTERS GENERATED A GREAT DEAL of excitement in the 1950s when they began to beat humans at checkers and to prove math theorems. In the 1960s the hope grew that scientists might soon be able to replicate the human brain in hardware and software and that "artificial intelligence" would soon match human performance on any task. In 1967 Marvin Minsky of the Massachusetts Institute of Technology, who died earlier this year, proclaimed that the challenge of AI would be solved within a generation.

IN BRIEF

Artificial intelligence started as a field of serious study in the mid-1950s. At the time, investigators expected to emulate human intelligence within the span of an academic career.

Hopes were dashed when it became clear that the algorithms and computing power of that period were simply not up to the task. Some skeptics even wrote off the endeavor as pure hubris.

A revival took place during the past few years as software patterned roughly after networks of neurons in the brain demonstrated that AI's early promise might yet be realized.

Deep learning—a technique that uses complex neural networks—has the ability to learn abstract concepts and already approaches human-level performance on some tasks.

That optimism, of course, turned out to be premature. Software designed to help physicians make better diagnoses and networks modeled after the human brain for recognizing the contents of photographs failed to live up to their initial hype. The algorithms of those early years lacked sophistication and needed more data than were available at the time. Computer processing was also too tepid to power machines that could perform the massive calculations needed to approximate something approaching the intricacies of human thought.

By the mid-2000s the dream of building machines with human-level intelligence had almost disappeared in the scientific community. At the time, even the term "AI" seemed to leave the domain of serious science. Scientists and writers describe the dashed hopes of the period from the 1970s until the mid-2000s as a series of "AI winters."

What a difference a decade makes. Beginning in 2005, AI's outlook changed spectacularly. That was when deep learning, an approach to building intelligent machines that drew inspiration from brain science, began to come into its own. In recent years deep learning has become a singular force propelling AI research forward. Major information technology companies are now pouring billions of dollars into its development.

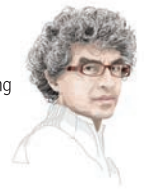
Deep learning refers to the simulation of networks of neurons that gradually "learn" to recognize images, understand speech or even make decisions on their own. The technique relies on so-called artificial neural networks—a core element of current AI research. Artificial neural networks do not mimic precisely how actual neurons work. Instead they are based on general mathematical principles that allow them to learn from examples to recognize people or objects in a photograph or to translate the world's major languages.

The technology of deep learning has transformed AI research, reviving lost ambitions for computer vision, speech recognition, natural-language processing and robotics. The first products rolled out in 2012 for understanding speech—you may be familiar with Google Now. And shortly afterward came applications for identifying the contents of an image, a feature now incorporated into the Google Photos search engine.

Anyone frustrated by clunky automated telephone menus can appreciate the dramatic advantages of using a better personal assistant on a smartphone. And for those who remember how poor object recognition was just a few years ago—software that might mistake an inanimate object for an animal—strides in computer vision have been incredible: we now have computers that, under certain conditions, can recognize a cat, a rock or faces in images almost as well as humans. AI software, in fact, has now become a familiar fixture in the lives of millions of smartphone users. Personally, I rarely type messages anymore. I often just speak to my phone, and sometimes it even answers back.

These advances have suddenly opened the door to further commercialization of the technology, and the excitement only continues to grow. Companies compete fiercely for talent, and Ph.D.s specializing in deep learning are a rare commodity that is in extremely high demand. Many university professors with expertise in this area—by some counts, the majority—have been pulled from academia to industry and furnished with well-appointed research facilities and ample compensation packages.

Working through the challenges of deep learning has led to stunning successes. The triumph of a neural network over top-ranked player Lee Se-dol at the game of Go received prominent headlines. Applications are already expanding to encompass other fields of human expertise—and it is not all games. A newly developed deep-learning algorithm is purported to diagnose heart failure from magnetic resonance imaging as well as a cardiologist.



Yoshua Bengio is a professor of computer science at the University of Montreal and one of the pioneers in developing the deep-learning methods that have sparked the current revival of artificial intelligence.

INTELLIGENCE, KNOWLEDGE AND LEARNING

WHY DID AI HIT so many roadblocks in previous decades? The reason is that most of the knowledge we have of the world around us is not formalized in written language as a set of explicit tasks—a necessity for writing any computer program. That is why we have not been able to directly program a computer to do many of the things that we humans do so easily—be it understanding speech, images or language or driving a car. Attempts to do so—organizing sets of facts in elaborate databases to imbue computers with a facsimile of intelligence—have met with scant success.

That is where deep learning comes in. It is part of the broader AI discipline known as machine learning, which is based on principles used to train intelligent computing systems—and to ultimately let machines teach themselves. One of these tenets relates to what a human or machine considers a "good" decision. For animals, evolutionary principles dictate decisions should be made that lead to behaviors that optimize chances of survival and reproduction. In human societies, a good decision might include social interactions that bring status or a sense of well-being. For a machine, such as a self-driving car, though, the quality of decision making depends on how closely the autonomous vehicle imitates the behaviors of competent human drivers.

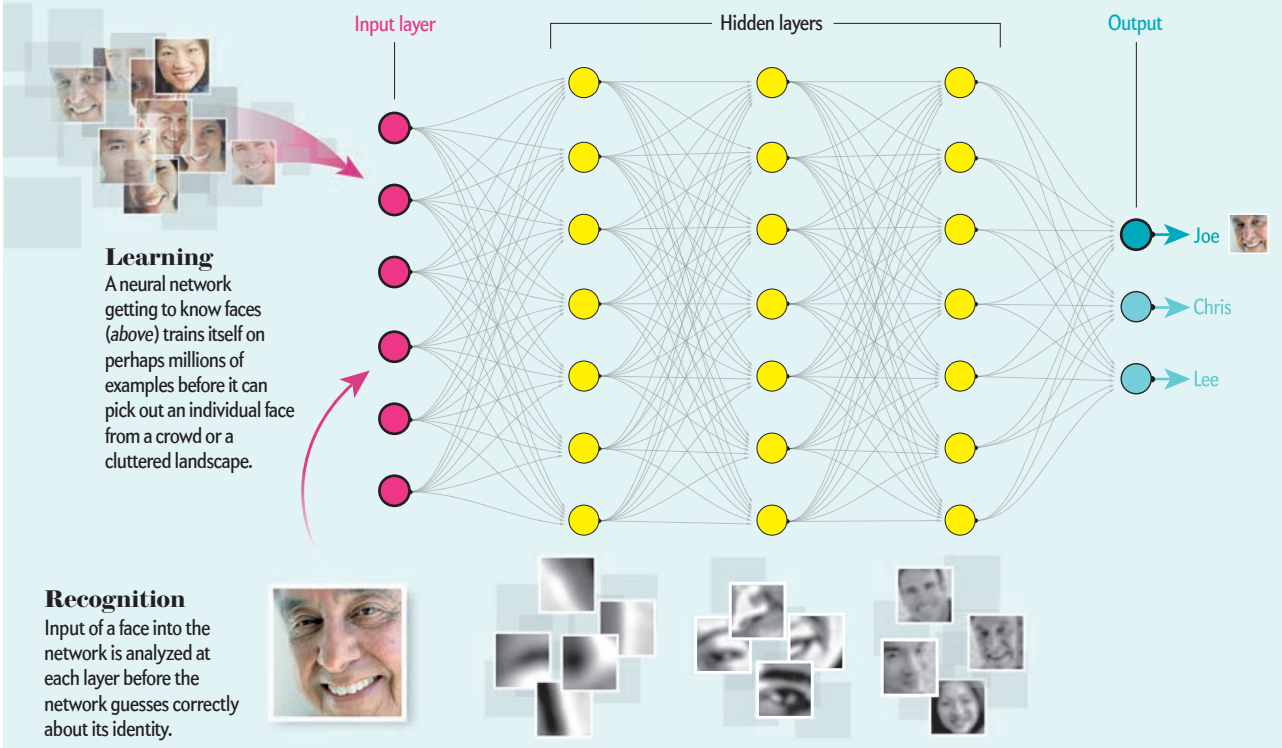
The knowledge needed to make a good decision in a particular context is not necessarily obvious in a way that can be translated into computer code. A mouse, for instance, has knowledge of its surroundings and an innate sense of where to sniff and how to move its legs, find food or mates, and avoid predators. No programmer would be capable of specifying a step-by-step set of instructions to produce these behaviors. Yet that knowledge is encoded in the rodent's brain.

Before creating computers that can train themselves, computer scientists needed to answer such fundamental questions as how humans acquire knowledge. Some knowledge is innate, but most is learned from experience. What we know intuitively cannot be turned into a clear sequence of steps for a computer to execute but can often be learned from examples and practice. Since the 1950s researchers have looked for and tried to refine general principles that allow animals or humans—or even machines, for that matter—to acquire knowledge through experience. Machine learning aims to establish procedures, called learning algorithms, that allow a machine to learn from examples presented to it.

Brainy Networks That Only Get Smarter

Connections from one neuron to the next in the brain's cortex have inspired the creation of algorithms that mimic these intricate links. A neural network can be trained to recognize a face by first training on countless images. Once it has "learned" to categorize a face (versus a hand, for instance) and to detect individual faces, the network uses that knowledge to identify faces it has seen before, even if the image of the person is slightly different from the one it was trained on.

To recognize a face, the network sets about the task of analyzing the individual pixels of an image presented to it at the input layer. Then, at the next layer, it chooses geometric shapes distinctive to a particular face. Moving up the hierarchy, a middle layer detects eyes, a mouth and other features before a composite full-face image is discerned at a higher layer. At the output layer, the network makes a "guess" about whether the face is that of Joe or rather that of Chris or Lee.



Learning
A neural network getting to know faces (*above*) trains itself on perhaps millions of examples before it can pick out an individual face from a crowd or a cluttered landscape.

Recognition
Input of a face into the network is analyzed at each layer before the network guesses correctly about its identity.

Each layer identifies progressively more complex features

PUNCHSTOCK (*faces*)

The science of machine learning is largely experimental because no universal learning algorithm exists—none can enable the computer to learn every task it is given well. Any knowledge-acquisition algorithm needs to be tested on learning tasks and data specific to the situation at hand, whether it is recognizing a sunset or translating English into Urdu. There is no way to prove that it will be consistently better across the board for any given situation than all other algorithms.

AI researchers have fashioned a formal mathematical description of this principle—the "no free lunch" theorem—that demonstrates that no algorithm exists to address every real-world learning situation. Yet human behavior apparently contradicts this theorem. We appear to hold in our head fairly general learning abilities that allow us to master a multitude of tasks for which evolution did not prepare our ancestors: playing chess, building bridges or doing research in AI.

These capabilities suggest that human intelligence exploits general assumptions about the world that might serve as inspiration for creating machines with a form of general intelligence. For just this reason, developers of artificial neural networks have adopted the brain as a rough model for designing intelligent systems.

The brain's main units of computation are cells called neurons. Each neuron sends a signal to other neurons through tiny gaps between the cells known as synaptic clefts. The propensity of a neuron to send a signal across the gap—and the amplitude of that signal—is referred to as synaptic strength. As a neuron "learns," its synaptic strength grows, and it is more likely, when stimulated by an electrical impulse, to send messages along to its neighbors.

Brain science influenced the emergence of artificial neural networks that used software or hardware to create virtual neurons. Early researchers in this subfield of AI, known as connectionism, postulated that neural networks would be able to learn

Graphic by Jen Christiansen

complex tasks by gradually altering the connections among neurons, so that patterns of neural activity would capture the content of its input, such as an image or a snippet of dialogue. As these networks would receive more examples, the learning process would continue by changing synaptic strengths among the connected neurons to achieve more accurate representations of, say, images of a sunset.

LESSONS ABOUT SUNSETS

THE CURRENT GENERATION of neural networks extends the pioneering work of connectionism. The networks gradually change numerical values for each synaptic connection, values representing the strength of that connection and thus how likely a neuron is to transmit a signal along to another neuron. An algorithm used by deep-learning networks changes these values ever so slightly each time it observes a new image. The values inch steadily closer toward ones that allow the neural network to make better predictions about the image's content.

For best results, current learning algorithms require close involvement by a human. Most of these algorithms use supervised learning in which each training example is accompanied by a human-crafted label about what is being learned—a picture of a sunset, say, is associated with a caption that says "sunset." In this instance, the goal of the supervised learning algorithm is to take a photograph as the input and produce, as an output, the name of a key object in the image. The mathematical process of transforming an input to an output is called a function. The numerical values, such as synaptic strengths, that produce this function correspond to a solution to the learning task.

Learning by rote to produce correct answers would be easy but somewhat useless. We want to teach the algorithm what a sunset is but then to have it recognize an image of any sunset, even one it has not been trained on. The ability to discern any sunset—in other words, to generalize learning beyond specific examples—is the main goal of any machine-learning algorithm. In fact, the quality of training of any network is evaluated by testing it using examples not previously seen. The difficulty of generalizing correctly to a new example arises because there is an almost infinite set of possible variations that still correspond to any category, such as a sunset.

To succeed in generalizing from having observed a multitude of examples, the learning algorithm used in deep-learning networks needs more than just the examples themselves. It also relies on hypotheses about the data and assumptions about what a possible solution to a particular problem might be. A typical hypothesis built into the software might postulate that if data inputs for a particular function are similar, the outputs should not radically change—altering a few pixels in an image of a cat should not usually transform the animal into a dog.

One type of neural network that incorporates hypotheses about images is called a convolutional neural network; it has become a key technology that has fueled the revival of AI. Convolutional neural networks used in deep learning have many layers of neurons organized in such a way as to make the output less sensitive to the main object in an image changing, such as when its position is moved slightly—a well-trained network may be able to recognize a face from different angles in separate photographs. The design of a convolutional network draws its inspiration from the multilayered structure of the visual cortex—the part of our brain that receives input from the eyes. The many layers of virtual neurons in a convolutional neural network are what makes a network "deep" and thus better able to learn about the world around it.

GOING DEEP

ON A PRACTICAL LEVEL, the advances that enabled deep learning came from specific innovations that emerged about 10 years ago, when interest in AI and neural networks had reached its lowest point in decades. A Canadian organization funded by the government and private donors, the Canadian Institute for Advanced Research (CIFAR), helped to rekindle the flame by sponsoring a program led by Geoffrey Hinton of the University of Toronto. The program also included Yann LeCun of New York University, Andrew Ng of Stanford University, Bruno Olshausen of the University of California, Berkeley, me and several others. Back then, negative attitudes toward this line of research made it difficult to publish and even to convince graduate students to work in this area, but a few of us had the strong sense that it was important to move ahead.

> The strong comeback for AI after a long and extended hiatus provides a lesson in the sociology of science, underscoring the need to put forward ideas that challenge the technological status quo.

Skepticism about neural networks at that time stemmed, in part, from the belief that training them was hopeless because of the challenges involved in optimizing how they behave. Optimization is a branch of mathematics that tries to find the configuration of a set of parameters to reach a mathematical objective. The parameters, in this case, are called synaptic weights and represent how strong a signal is being sent from one neuron to another.

The objective is to make predictions with the minimum number of errors. When the relation between parameters and an objective is simple enough—more precisely when the objective is a convex function of the parameters—the parameters can be gradually adjusted. This continues until they get as close as

possible to the values that produce the best possible choice, known as a global minimum—which corresponds to the lowest possible average prediction error made by the network.

In general, however, training a neural network is not so simple—and requires what is called a nonconvex optimization. This type of optimization poses a much greater challenge—and many researchers believed that the hurdle was insurmountable. The learning algorithm can get stuck in what is called a local minimum, in which it is unable to reduce the prediction error of the neural network by adjusting parameters slightly.

Only in the past year was the myth dispelled that neural networks were hard to train because of local minima. We found in our research that when a neural network is sufficiently large, the local minima problem is greatly reduced. Most local minima actually correspond to having learned knowledge at a level that almost matches the optimal value of the global minimum.

Although the theoretical problems of optimization could, in theory, be solved, building large networks with more than two or three layers had often failed. Beginning in 2005, CIFAR-supported efforts achieved breakthroughs that overcame these barriers. In 2006 we managed to train deeper neural networks, using a technique that proceeded layer by layer.

Later, in 2011, we found a better way to train even deeper networks—ones with more layers of virtual neurons—by altering the computations performed by each of these processing units, making them more like what biological neurons actually compute. We also discovered that injecting random noise into the signals transmitted among neurons during training, similar to what happens in the brain, made them better able to learn to correctly identify an image or sound.

Two crucial factors aided the success of deep-learning techniques. An immediate 10-fold increase in computing speed, thanks to the graphics-processing units initially designed for video games, allowed larger networks to be trained in a reasonable amount of time. Also fueling deep learning's growth was the availability of huge labeled data sets for which a learning algorithm can identify the correct answer—"cat," for example, when inspecting an image in which a cat is just one element.

Another reason for deep learning's recent success is its ability to learn to perform a sequence of computations that construct or analyze, step by step, an image, a sound or other data. The depth of the network is the number of such steps. Many visual- or auditory-recognition tasks in which AI excels require the many layers of a deep network. In recent theoretical and experimental studies, in fact, we have shown that carrying out some of these mathematical operations cannot be accomplished efficiently without sufficiently deep networks.

Each layer in a deep neural network transforms its input and produces an output that is sent to the next layer. The network represents more abstract concepts at its deeper layers [*see box on page 49*], which are more remote from the initial raw sensory input. Experiments show that artificial neurons in deeper layers in the network tend to correspond to more abstract semantic concepts: a visual object such as a desk, for instance. Recognition of the image of the desk might emerge from the processing of neurons at a deeper layer even though the concept of "desk" was not among the category labels on which the network was trained. And the concept of a desk might itself only be an intermediate step toward creating a still more abstract concept at a still higher layer that might be categorized by the network as an "office scene."

BEYOND PATTERN RECOGNITION

UNTIL RECENTLY, artificial neural networks distinguished themselves in large part for their ability to carry out tasks such as recognizing patterns in static images. But another type of neural network is also making its mark—specifically, for events that unfold over time. Recurrent neural networks have demonstrated the capacity to correctly perform a sequence of computations, typically for speech, video and other data. Sequential data are made up of units—whether a phoneme or a whole word—that follow one another sequentially. The way recurrent neural networks process their inputs bears a resemblance to how the brain works. Signals that course among neurons change constantly as inputs from the senses are processed. This internal neural state changes in a way that depends on the current input to the brain from its surroundings before issuing a sequence of commands that result in body movements directed at achieving a specific goal.

Recurrent networks can predict what the next word in a sentence will be, and this can be used to generate new sequences of words, one at a time. They can also take on more sophisticated tasks. After "reading" all the words in a sentence, the network can guess at the meaning of the entire sentence. A separate recurrent network can then use the semantic processing of the first network to translate the sentence into another language.

Research on recurrent neural networks had its own lull in the late 1990s and early 2000s. My theoretical work suggested that they would run into difficulty learning to retrieve information from the far past—the earliest elements in the sequence being processed. Think of trying to recite the words from the first sentences of a book verbatim when you have just reached the last page. But several advances have lessened some of these problems by enabling such networks to learn to store information so that it persists for an extended time. The neural networks can use a computer's temporary memory to process multiple, dispersed pieces of information, such as ideas contained in different sentences spread across a document.

The strong comeback for deep neural networks after the long AI winter is not just a technological triumph. It also provides a lesson in the sociology of science. In particular, it underscores the need to support ideas that challenge the technological status quo and to encourage a diverse research portfolio that backs disciplines that temporarily fall out of favor. SA

MORE TO EXPLORE

ImageNet Classification with Deep Convolutional Neural Networks. Alex Krizhevsky et al. Presented at the 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), Stateline, Nev., December 3–8, 2012.

Representation Learning: A Review and New Perspectives. Y. Bengio et al. in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 35, No. 8, pages 1798–1828; August 2013.

Deep Learning. Yann LeCun et al. in *Nature,* Vol. 521, pages 436–444; May 28, 2015.

FROM OUR ARCHIVES

When Computers Surpass Us. Christof Koch; Consciousness Redux, *Scientific American Mind,* September/October 2015.