Harvard Business Review

REPRINT H03F8N
PUBLISHED ON HBR.ORG
JANUARY 30, 2017

ARTICLE
TECHNOLOGY

Deep Learning Will Radically Change the Ways We Interact with Technology

by Aditya Singh

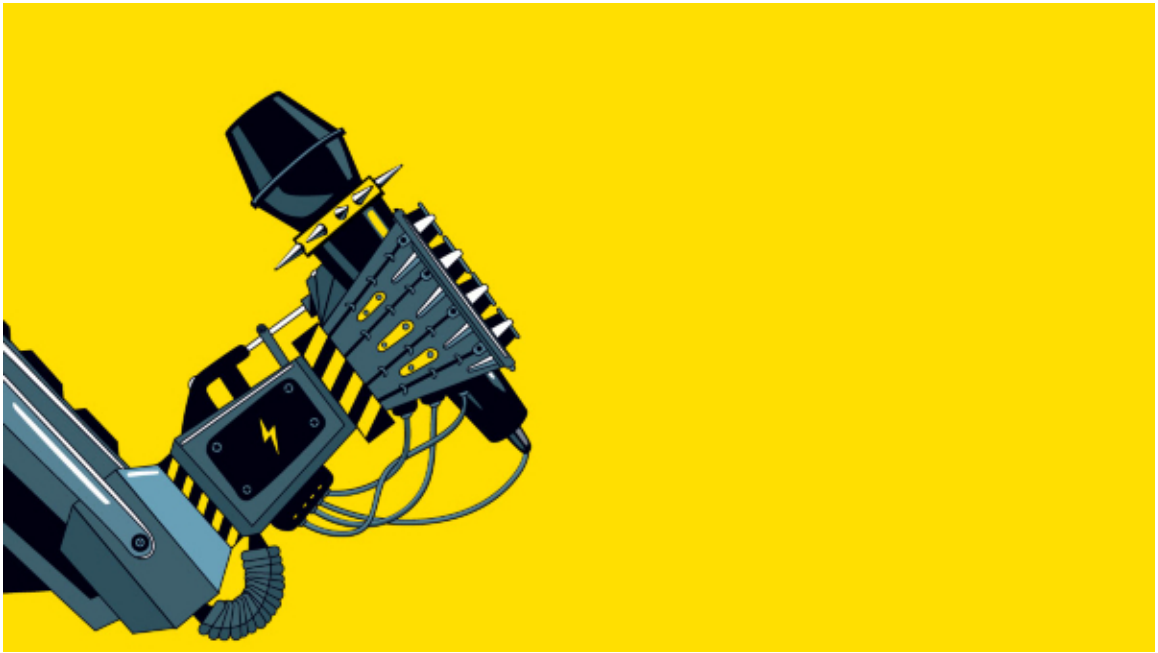
Harvard Business Review

TECHNOLOGY

Deep Learning Will Radically Change the Ways We Interact with Technology

by Aditya Singh

JANUARY 30, 2017



Even though heat and sound are both forms of energy, when you were a kid, you probably didn't need to be told not to speak in thermal convection. And each time your children come across a stray animal, they likely don't have to self-consciously rehearse a subroutine of zoological attributes to decide whether it's a cat or a dog. Human beings come pre-loaded with the cognitive gear to simply

perceive these distinctions. The differences appear so obvious, and knowing the differences comes so naturally to us, that we refer to it as common sense. Computers, in contrast, need step-by-step handholding—in the form of deterministic algorithms—to render even the most basic of judgments. Despite decades of unbroken gains in speed and processing capacity, machines can't do what the average toddler does without even trying. That is—until now.

Over the last half-dozen years, deep learning, a branch of artificial intelligence inspired by the structure of the human brain, has made enormous strides in giving machines the ability to intuit the physical world. At Facebook's AI lab, they've built a deep learning system capable of answering simple questions to which it had never previously been exposed. The Echo, Amazon's smart speaker, uses deep learning techniques. Three years ago, Microsoft's chief research officer impressed attendees at a lecture in China with a demonstration of deep learning speech software that translated his spoken English into Chinese, then instantly delivered the translation using a simulation of his voice speaking Mandarin—with an error rate of just 7%. It now uses the technology to improve voice search on Windows mobile and Bing.

The most powerful tech companies in the world have been quietly deploying deep learning to improve their products and services, and none has invested more than Google. It has "bet the company" on AI, says the New York Times, committing huge resources and scooping up many of the leading researchers in the field. And its efforts have borne fruit. A few years ago, a Google deep learning network was shown 10 million unlabeled images from YouTube, and proved to be nearly twice as accurate at identifying the objects in the images (cats, human faces, flowers, various species of fish, and thousands of others) as any previous method. When Google deployed deep learning on its Android voice search, errors dropped by 25% overnight. At the beginning of this year, another Google deep learning system defeated one of the best players of Go—the world's most complex board game.

This is only the beginning. I believe that over the next few years start-ups and the usual big tech suspects will use deep learning to upgrade a wide suite of existing applications, and to create new products and services. Entirely new business lines and markets will spring up, which will, in turn, give rise to still more innovation. Deep learning systems will become easier to use and more widely available. And I predict that deep learning will change the way people interact with technology as radically as operating systems transformed ordinary people's access to computers.

Deep Learning

Historically, computers performed tasks by being programmed with deterministic algorithms, which detailed every step that had to be taken. This worked well in many situations, from performing elaborate calculations to defeating chess grandmasters. But it hasn't worked as well in situations where providing an explicit algorithm wasn't possible—such as recognizing faces or emotions, or answering novel questions.

Trying to approach those challenges by hand-coding the myriad attributes of a face or phoneme was too labor-intensive, and left machines unable to process data that didn't fit within the explicit

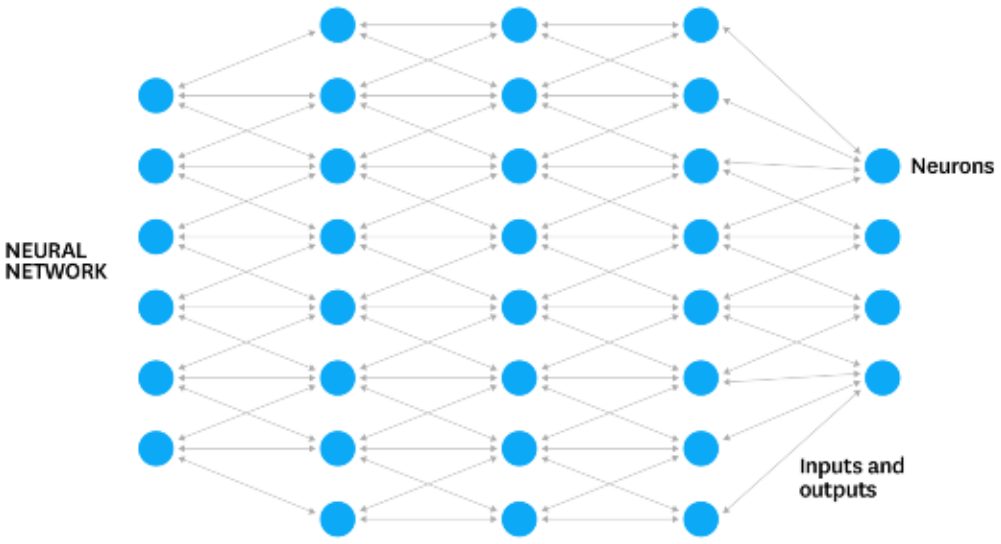
parameters provided by the programmers. Think of the difference between modern voice-assistants like Siri or Alexa, which allow you to ask for things in various ways using natural language, vs. automated phone menu systems, which only perform if you use the specific set of non-negotiable words that they were programmed to understand. By contrast, deep learning-based systems make sense of data for themselves, without the need of an explicit algorithm. Loosely inspired by the human brain, these machines learn, in a very real sense, from their experience. And some are now about as good at object and speech recognition as people.

So how does deep learning work?

Deep learning systems are modeled after the neural networks in the neocortex of the human brain, where higher-level cognition occurs. In the brain, a neuron is a cell that transmits electrical or chemical information. When connected with other neurons, it forms a neural network. In machines, the neurons are virtual—basically bits of code running statistical regressions. String enough of these virtual neurons together and you get a virtual neural network. Think of every neuron in the network below as a simple statistical model: it takes in some inputs, and it passes along some output.

Deep Learning Consists of Neural Networks

These computational models are loosely inspired by the human brain, where neurons take input and pass along outputs.



SOURCE FOUNDATION CAPITAL

© HBR.ORG

For a neural network to be useful, though, it requires training. To train a neural network, a set of virtual neurons are mapped out and assigned a random numerical "weight," which determines how the neurons respond to new data (digitized objects or sounds). Like in any statistical or machine learning, the machine initially gets to see the correct answers, too. So if the network doesn't accurately identify the input – doesn't see a face in an image, for example — then the system adjusts

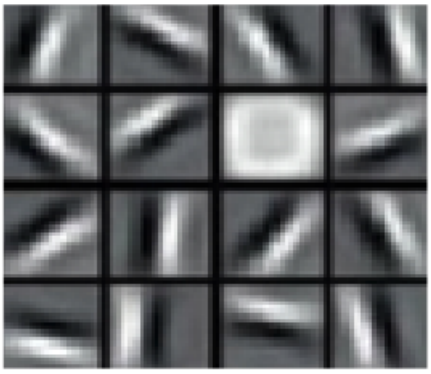
the weights—i.e., how much attention each neuron paid to the data—in order to produce the right answer. Eventually, after sufficient training, the neural network will consistently recognize the correct patterns in speech or images.

The idea of artificial neurons has been around for at least 60 years, when, in the 1950s, Frank Rosenblatt built a "perceptron" made of motors, dials, and light detectors, which he successfully trained to tell the difference between basic shapes. But early neural networks were extremely limited in the number of neurons they could simulate, which meant they couldn't recognize complex patterns. Three developments in the last decade made deep learning viable.

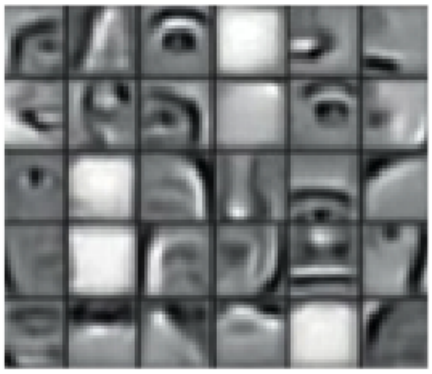
First, Geoffrey Hinton and other researchers at the University of Toronto developed a breakthrough method for software neurons to teach themselves by layering their training. (Hinton now splits his time between the University of Toronto and Google.) A first layer of neurons will learn how to distinguish basic features, say, an edge or a contour, by being blasted with millions of data points. Once the layer learns how to recognize these things accurately, it gets fed to the next layer, which trains itself to identify more complex features, say, a nose or an ear. Then that layer gets fed to another layer, which trains itself to recognize still greater levels of abstraction, and so on, layer after layer—hence the "deep" in deep learning—until the system can reliably recognize very complex phenomenon, like a human face.

Deep Learning Relies on Multiple "Layers" of Training

In this case, the first layer recognizes edges, the second recognizes facial features like a nose or an ear, until eventually the final layer recognizes full faces.



1. EDGES



2. FEATURES

3. FACES

4. FULL FACE

SOURCE "UNSUPERVISED LEARNING OF HIERARCHICAL REPRESENTATIONS WITH CONVOLUTIONAL DEEP BELIEF NETWORKS," BY HONGLAK LEE ET AL, 2011

© HBR.ORG

The second development responsible for recent advancements in AI is the sheer amount of data that is now available. Rapid digitization has resulted in the production of large-scale data, and that data is oxygen for training deep learning systems. Children can pick something up after being shown how to do it just a few times. AI-powered machines, however, need to be exposed to countless examples. Deep learning is essentially a brute-force process for teaching machines how a thing is done or what a thing is. Show a deep learning neural network 19 million pictures of cats and probabilities emerge, inclinations are ruled out, and the software neurons eventually figure out what statistically significant factors equate to feline. It learns how to spot a cat. That's why Big Data is so important—without it, deep learning just doesn't work.

Finally, a team at Stanford led by Andrew Ng (now at Baidu) made a breakthrough when they realized that graphics processing unit chips, or GPUs, which were invented for the visual processing demands of video games, could be repurposed for deep learning. Until recently, typical computer chips could only process one event at a time, but GPUs were designed for *parallel* computing. Using these chips to run neural networks, with their millions of connections, in parallel sped up the training and abilities of deep learning systems by several orders of magnitude. It made it possible for a machine to learn in a day something that had previously taken many weeks.

The most advanced deep learning networks today are made up of millions of simulated neurons, with billions of connections between them, and can be trained through unsupervised learning. It is the most effective practical application of artificial intelligence that's yet been devised. For some tasks, the best deep learning systems are pattern recognizers on par with people. And the technology is moving aggressively from the research lab into industry.

Deep Learning OS 1.0

As impressive as the gains from deep learning have been already, these are early days. If I analogize it to the personal computer, deep learning is in the green-and-black-DOS-screen stage of its evolution. A great deal of time and effort, at present, is being spent doing *for* deep learning—cleaning, labelling, and interpreting data, for example—rather than doing *with* deep learning. But in the next couple of years, start-ups and established companies will begin releasing commercial solutions for building production-ready deep learning applications. Making use of open-source frameworks such as TensorFlow, these solutions will dramatically reduce the effort, time, and costs of creating complex deep learning systems. Together they will constitute the building blocks of a deep learning operating system.

A deep learning operating system will permit the widespread adoption of practical AI. In the same way that Windows and Mac OS allowed regular consumers to use computers and SaaS gave them access to the cloud, tech companies in the next few years will democratize deep learning. Eventually, a deep learning OS will allow people who aren't computer scientists or natural language processing researchers to use deep learning to solve real-life problems, like detecting diseases instead of identifying cats.

The first new companies making up the deep learning operating system will be working on solutions in data, software, hardware.

Data. Getting good quality large scale data is the biggest barrier to adopting deep learning. But both service shops and software platforms will arise to deal with the data problem. Companies are already creating internal intelligent platforms that assist humans to label data quickly. Future data labeling platforms will be embedded in the design of the application, so that the data created by using a product will be captured for training purposes. And there will be new service-based companies that will outsource labeling to low-cost countries, as well as create labeled data through synthetic means.

Software. There are two main areas here where I see innovation happening:

1) *The design and programming of neural networks*. Different deep learning architectures, such as CNNs and RNNs, support different types of applications (image, text, etc.). Some use a combination of neural network architectures. As for training, many applications will use a combination of machine learning algorithms, deep learning, reinforcement learning, or unsupervised learning for solving different sub-parts of the application. I predict that someone will build a machine learning design engine solution, which will examine an application, training data set, infrastructure resources, and so on, and recommend the right architecture and algorithms to be used.

2) *A marketplace of reusable neural network modules*. As described above, different layers in a neural network learn different concepts and then build on each other. This architecture naturally creates opportunity to share and reuse trained neural networks. A layer of virtual neurons that's been trained to identify an edge, on its way up to recognizing the face of cat, could also be repurposed as the base layer for recognizing the face of a person. Already, Tensorflow, the most popular deep learning framework, supports reusing an entire subgraph component. Soon, the community of machine learning experts contributing open source modules will create the potential for deep learning versions of GitHub and StackOverflow.

Hardware. Finding the optimal mix of GPUs, CPUs, cloud resources; determining the level of parallelization; and performing cost analyses are complex decisions for developers. This creates an opportunity for platform and service-based companies to recommend the right infrastructure for training tasks. Additionally, there will be companies that provide infrastructure services—such as orchestration, scale-out, management, and load balancing—on specialized hardware for deep learning. Moreover, I expect incumbents as well as start-ups to launch their own deep learning-optimized chips.

These are just some of the possibilities. I'm certain there are many more lurking in other entrepreneurial minds, because the promise of this technology is immense. We are beginning to build machines that can learn for themselves and that have some semblance of sensible judgment.

Palak Dalal (HBS '17) contributed research and analysis to this piece.

Aditya Singh is a partner at Foundation Capital.