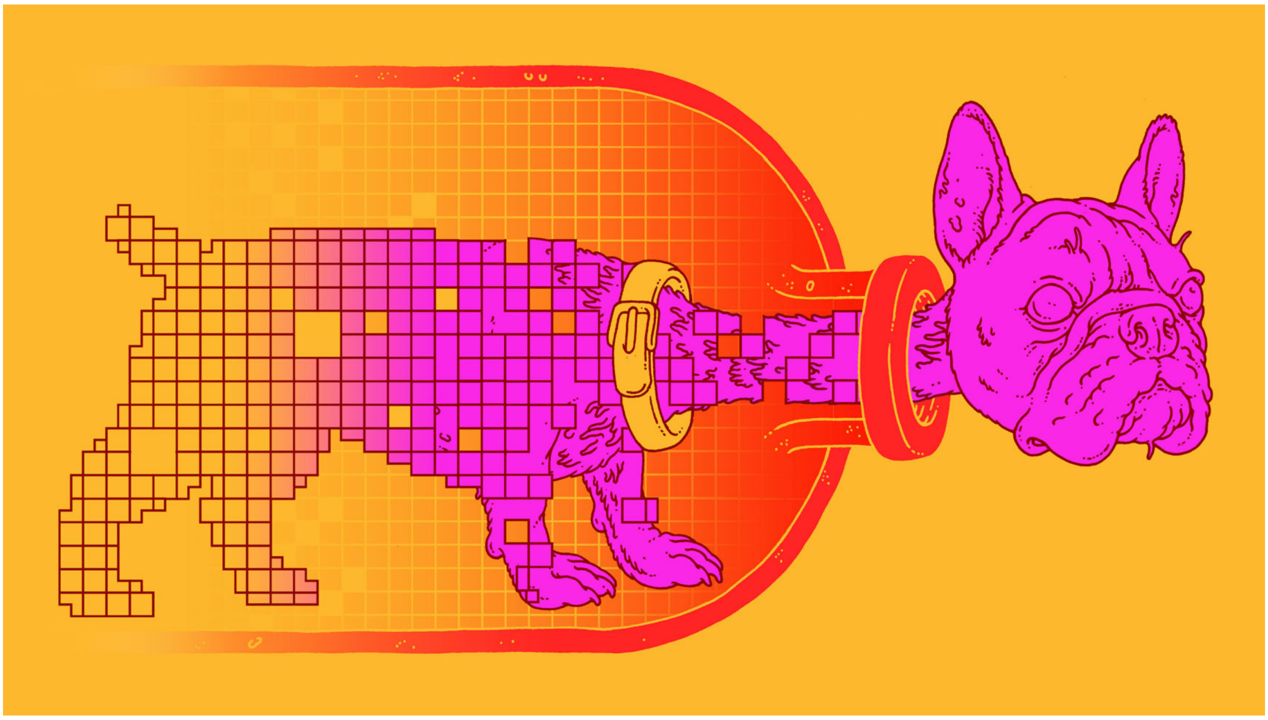
Quanta magazine

New Theory Cracks Open the Black Box of Deep Learning

A new idea called the "information bottleneck" is helping to explain the puzzling success of today's artificial-intelligence algorithms — and might also explain how human brains learn.

By Natalie Wolchover



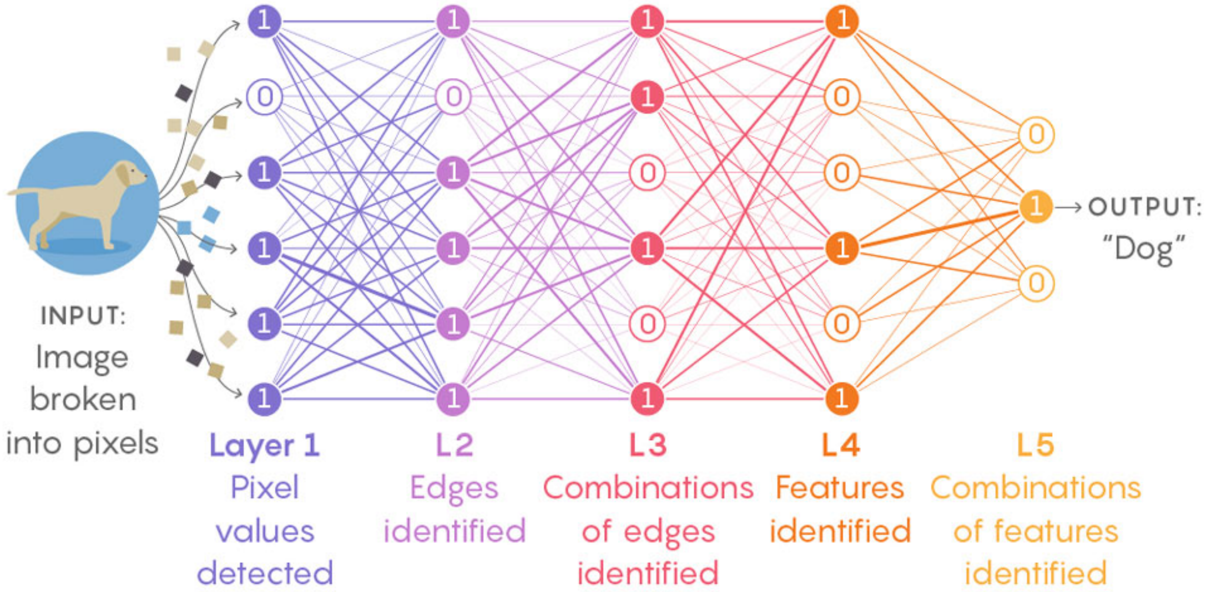
Eric Nyquist for Quanta Magazine

Even as machines known as "deep neural networks" have learned to converse, drive cars, beat video games and Go champions, dream, paint pictures and help make scientific discoveries, they have also confounded their human creators, who never expected so-called "deep-learning" algorithms to work so well. No underlying principle has guided the design of these learning systems, other than vague inspiration drawn from the architecture of the brain (and no one really understands how that operates either).

Like a brain, a deep neural network has layers of neurons — artificial ones that are figments of computer memory. When a neuron fires, it sends signals to connected neurons in the layer above. During deep learning, connections in the network are strengthened or weakened as needed to make the system better at sending signals from input data — the pixels of a photo of a dog, for instance — up through the layers to neurons associated with the right high-level concepts, such as “dog.” After a deep neural network has “learned” from thousands of sample dog photos, it can identify dogs in new photos as accurately as people can. The magic leap from special cases to general concepts during learning gives deep neural networks their power, just as it underlies human reasoning, creativity and the other faculties collectively termed “intelligence.” Experts wonder what it is about deep learning that enables generalization — and to what extent brains apprehend reality in the same way.

Learning From Experience

Deep neural networks learn by adjusting the strengths of their connections to better convey input signals through multiple layers to neurons associated with the right general concepts.



When data is fed into a network, each artificial neuron that fires (labeled “1”) transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. The process filters out noise and retains only the most relevant features.

Lucy Reading-Ikkanda/Quanta Magazine

Last month, a YouTube video of a conference talk in Berlin, shared widely among artificial-intelligence researchers, offered a possible answer. In the talk, Naftali Tishby, a computer scientist and neuroscientist from the Hebrew University of Jerusalem, presented evidence in support of a new theory explaining how deep learning works. Tishby argues that deep neural networks learn according to a procedure called the "information bottleneck," which he and two collaborators first described in purely theoretical terms in 1999. The idea is that a network rids noisy input data of extraneous details as if by squeezing the information through a bottleneck, retaining only the features most relevant to general concepts. Striking new computer experiments by Tishby and his student Ravid Shwartz-Ziv reveal how this squeezing procedure happens during deep learning, at least in the cases they studied.

Tishby's findings have the AI community buzzing. "I believe that the information bottleneck idea could be very important in future deep neural network research," said Alex Alemi of Google Research, who has already developed new approximation methods for applying an information bottleneck analysis to large deep neural networks. The bottleneck could serve "not only as a theoretical tool for understanding why our neural networks work as well as they do currently, but also as a tool for constructing new objectives and architectures of networks," Alemi said.

Some researchers remain skeptical that the theory fully accounts for the success of deep learning, but Kyle Cranmer, a particle physicist at New York University who uses machine learning to analyze particle collisions at the Large Hadron Collider, said that as a general principle of learning, it "somehow smells right."

Geoffrey Hinton, a pioneer of deep learning who works at Google and the University of Toronto, emailed Tishby after watching his Berlin talk. "It's extremely interesting," Hinton wrote. "I have to listen to it another 10,000 times to really understand it, but it's very rare nowadays to hear a talk with a really original idea in it that may be the answer to a really major puzzle."

According to Tishby, who views the information bottleneck as a fundamental principle behind learning, whether you're an algorithm, a housefly, a conscious being, or a physics calculation of emergent behavior, that long-awaited answer "is that the most important part of learning is actually forgetting."

The Bottleneck

Tishby began contemplating the information bottleneck around the time that other researchers were first mulling over deep neural networks, though neither concept had been named yet. It was the 1980s, and Tishby was thinking about how good humans are at speech recognition — a major challenge for AI at the time. Tishby realized that the crux of the issue was the question of relevance: What are the most relevant features of a spoken word, and how do we tease these out from the variables that accompany them, such as accents, mumbling and intonation? In general, when we face the sea of data that is reality, which signals do we keep?

Miriam Alster, Flash 90. ELSC Art and Brain Week 2016

Naftali Tishby, a professor of computer science at the Hebrew University of Jerusalem.

"This notion of relevant information was mentioned many times in history but never formulated correctly," Tishby said in an interview last month. "For many years people thought information theory wasn't the right way to think about relevance, starting with misconceptions that go all the way to Shannon himself."

Claude Shannon, the founder of information theory, in a sense liberated the study of information starting in the 1940s by allowing it to be considered in the abstract — as 1s and 0s with purely mathematical meaning. Shannon took the view that, as Tishby put it, "information is not about semantics." But, Tishby argued, this isn't true. Using information theory, he realized, "you can define 'relevant' in a precise sense."

Imagine X is a complex data set, like the pixels of a dog photo, and Y is a simpler variable represented by those data, like the word "dog." You can capture all the "relevant" information in X about Y by compressing X as much as you can without losing the ability to predict Y. In their 1999 paper, Tishby and co-authors Fernando Pereira, now at Google, and William Bialek, now at Princeton University, formulated this as a mathematical optimization problem. It was a fundamental idea with no killer application.

"I've been thinking along these lines in various contexts for 30 years," Tishby said. "My only luck was that deep neural networks became so important."

Eyeballs on Faces on People on Scenes

Though the concept behind deep neural networks had been kicked around for decades, their performance in tasks like speech and image recognition only took off in the early 2010s, due to improved training regimens and more powerful computer processors. Tishby recognized their potential connection to the information bottleneck principle in 2014 after reading a surprising paper by the physicists David Schwab and Pankaj Mehta.

The duo discovered that a deep-learning algorithm invented by Hinton called the “deep belief net” works, in a particular case, exactly like renormalization, a technique used in physics to zoom out on a physical system by coarse-graining over its details and calculating its overall state. When Schwab and Mehta applied the deep belief net to a model of a magnet at its “critical point,” where the system is fractal, or self-similar at every scale, they found that the network automatically used the renormalization-like procedure to discover the model’s state. It was a stunning indication that, as the biophysicist Ilya Nemenman said at the time, “extracting relevant features in the context of statistical physics and extracting relevant features in the context of deep learning are not just similar words, they are one and the same.”

The only problem is that, in general, the real world isn’t fractal. “The natural world is not ears on ears on ears on ears; it’s eyeballs on faces on people on scenes,” Cranmer said. “So I wouldn’t say [the renormalization procedure] is why deep learning on natural images is working so well.” But Tishby, who at the time was undergoing chemotherapy for pancreatic cancer, realized that both deep learning and the coarse-graining procedure could be encompassed by a broader idea. “Thinking about science and about the role of my old ideas was an important part of my healing and recovery,” he said.

Yifat Yogev (left); courtesy of Ravid Shwartz-Ziv (right)

Noga Zaslavsky, left, and Ravid Shwartz-Ziv helped develop the information bottleneck theory of deep learning as graduate students of Naftali Tishby's.

In 2015, he and his student Noga Zaslavsky hypothesized that deep learning is an information bottleneck procedure that compresses noisy data as much as possible while preserving information about what the data represent. Tishby and Shwartz-Ziv's new experiments with deep neural networks reveal how the bottleneck procedure actually plays out. In one case, the researchers used small networks that could be trained to label input data with a 1 or 0 (think "dog" or "no dog") and gave their 282 neural connections random initial strengths. They then tracked what happened as the networks engaged in deep learning with 3,000 sample input data sets.

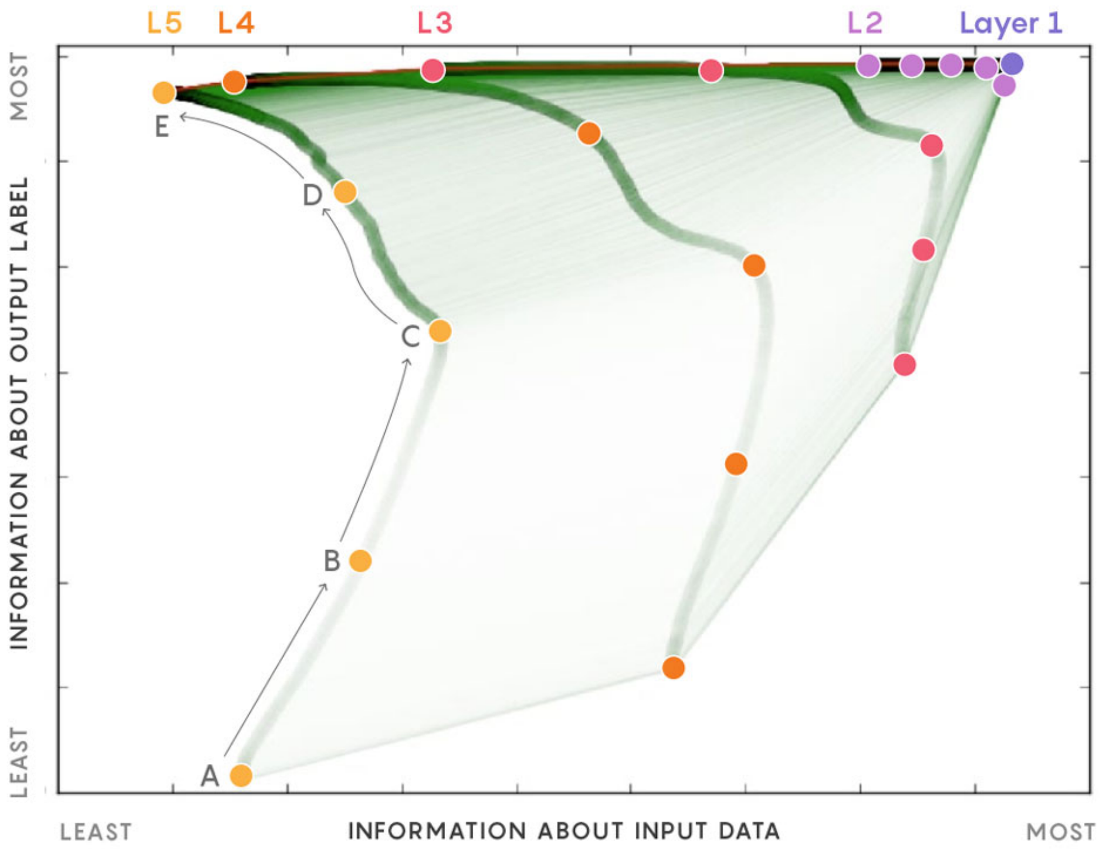
The basic algorithm used in the majority of deep-learning procedures to tweak neural connections in response to data is called "stochastic gradient descent": Each time the training data are fed into the network, a cascade of firing activity sweeps upward through the layers of artificial neurons. When the signal reaches the top layer, the final firing pattern can be compared to the correct label for the image — 1 or 0, "dog" or "no dog." Any differences between this firing pattern and the correct pattern are "back-propagated" down the layers, meaning that, like a teacher correcting an exam, the algorithm strengthens or weakens each connection to make the network layer better at producing the correct output signal. Over the course of training, common patterns in the training data become reflected in the strengths of the connections, and the network becomes expert at correctly labeling the data, such as by recognizing a dog, a word, or a 1.

In their experiments, Tishby and Shwartz-Ziv tracked how much information each layer of a deep neural network retained about the input data and how much information each one retained about the output label. The scientists found that, layer by layer, the networks converged to the information bottleneck theoretical bound: a theoretical limit derived in Tishby, Pereira and Bialek's original paper that represents the absolute best the system can do at extracting relevant information. At the bound, the network has compressed the input as much as possible without sacrificing the ability to accurately predict its label.

Tishby and Shwartz-Ziv also made the intriguing discovery that deep learning proceeds in two phases: a short "fitting" phase, during which the network learns to label its training data, and a much longer "compression" phase, during which it becomes good at generalization, as measured by its performance at labeling new test data.

Inside Deep Learning

New experiments reveal how deep neural networks evolve as they learn.



A **INITIAL STATE:** Neurons in Layer 1 encode everything about the input data, including all information about its label. Neurons in the highest layers are in a nearly random state bearing little to no relationship to the data or its label.

B **FITTING PHASE:** As deep learning begins, neurons in higher layers gain information about the input and get better at fitting labels to it.

C **PHASE CHANGE:** The layers suddenly shift gears and start to "forget" information about the input.

D **COMPRESSION PHASE:** Higher layers compress their representation of the input data, keeping what is most relevant to the output label. They get better at predicting the label.

E **FINAL STATE:** The last layer achieves an optimal balance of accuracy and compression, retaining only what is needed to predict the label.

Lucy Reading-Ikkanda/Quanta Magazine; adapted from arXiv:1703.00810 [cs.LG].

As a deep neural network tweaks its connections by stochastic gradient descent, at first the number of bits it stores about the input data stays roughly constant or increases slightly, as connections adjust to encode patterns in the input and the network gets good at fitting labels to it. Some experts have compared this phase to memorization.

Then learning switches to the compression phase. The network starts to shed information about the input data, keeping track of only the strongest features — those correlations that are most relevant to the output label. This happens because, in each iteration of stochastic gradient descent, more or less accidental correlations in the training data tell the network to do different things, dialing the strengths of its neural connections up and down in a random walk. This randomization is effectively the same as compressing the system's representation of the input data. As an example, some photos of dogs might have houses in the background, while others don't. As a network cycles through these training photos, it might "forget" the correlation between houses and dogs in some photos as other photos counteract it. It's this forgetting of specifics, Tishby and Shwartz-Ziv argue, that enables the system to form general concepts. Indeed, their experiments revealed that deep neural networks ramp up their generalization performance during the compression phase, becoming better at labeling test data. (A deep neural network trained to recognize dogs in photos might be tested on new photos that may or may not include dogs, for instance.)

It remains to be seen whether the information bottleneck governs all deep-learning regimes, or whether there are other routes to generalization besides compression. Some AI experts see Tishby's idea as one of many important theoretical insights about deep learning to have emerged recently. Andrew Saxe, an AI researcher and theoretical neuroscientist at Harvard University, noted that certain very large deep neural networks don't seem to need a drawn-out compression phase in order to generalize well. Instead, researchers program in something called early stopping, which cuts training short to prevent the network from encoding too many correlations in the first place.

Tishby argues that the network models analyzed by Saxe and his colleagues differ from standard deep neural network architectures, but that nonetheless, the information bottleneck theoretical bound defines these networks' generalization performance better than other methods. Questions about whether the bottleneck holds up for larger neural networks are partly addressed by Tishby and Shwartz-Ziv's most recent experiments, not included in their preliminary paper, in which they train much larger, 330,000-connection-deep neural networks to recognize handwritten digits in the 60,000-image Modified National Institute of Standards and Technology database, a well-known benchmark for gauging the performance of deep-learning algorithms. The scientists saw the same convergence of the networks to the information bottleneck theoretical bound; they also observed the two distinct phases of deep learning, separated by an even sharper transition than in the smaller networks. "I'm completely convinced now that this is a general phenomenon," Tishby said.

Humans and Machines

The mystery of how brains sift signals from our senses and elevate them to the level of our conscious awareness drove much of the early interest in deep neural networks among AI pioneers, who hoped to reverse-engineer the brain's learning rules. AI practitioners have since largely abandoned that path in the mad dash for technological progress, instead slapping on bells and whistles that boost performance with little regard for biological plausibility. Still, as their thinking machines achieve ever greater feats — even stoking fears that AI could someday pose an existential threat — many researchers hope these explorations will uncover general insights about learning and intelligence.

Brenden Lake, an assistant professor of psychology and data science at New York University who studies similarities and differences in how humans and machines learn, said that Tishby's findings represent "an important step towards opening the black box of neural networks," but he stressed that the brain represents a much bigger, blacker black box. Our adult brains, which boast several hundred trillion connections between 86 billion neurons, in all likelihood employ a bag of tricks to enhance generalization, going beyond the basic image- and sound-recognition learning procedures that occur during infancy and that may in many ways resemble deep learning.

For instance, Lake said the fitting and compression phases that Tishby identified don't seem to have analogues in the way children learn handwritten characters, which he studies. Children don't need to see thousands of examples of a character and compress their mental representation over an extended period of time before they're able to recognize other instances of that letter and write it themselves. In fact, they can learn from a single example. Lake and his colleagues' models suggest the brain may deconstruct the new letter into a series of strokes — previously existing mental constructs — allowing the conception of the letter to be tacked onto an edifice of prior knowledge. "Rather than thinking of an image of a letter as a pattern of pixels and learning the concept as mapping those features" as in standard machine-learning algorithms, Lake explained, "instead I aim to build a simple causal model of the letter," a shorter path to generalization.

Such brainy ideas might hold lessons for the AI community, furthering the back-and-forth between the two fields. Tishby believes his information bottleneck theory will ultimately prove useful in both disciplines, even if it takes a more general form in human learning than in AI. One immediate insight that can be gleaned from the theory is a better understanding of which kinds of problems can be solved by real and artificial neural networks. "It gives a complete characterization of the problems that can be learned," Tishby said. These are "problems where I can wipe out noise in the input without hurting my ability to classify. This is natural vision problems, speech recognition. These are also precisely the problems our brain can cope with."

Meanwhile, both real and artificial neural networks stumble on problems in which every detail matters and minute differences can throw off the whole result. Most people can't quickly multiply two large numbers in their heads, for instance. "We have a long class of problems like this, logical problems that are very sensitive to changes in one variable," Tishby said. "Classifiability, discrete problems, cryptographic problems. I don't think deep learning will ever help me break cryptographic codes."

Generalizing — traversing the information bottleneck, perhaps — means leaving some details behind. This isn't so good for doing algebra on the fly, but that's not a brain's main business. We're looking for familiar faces in the crowd, order in chaos, salient signals in a noisy world.