A Common Logic to Seeing Cats and Cosmos

By Natalie Wolchover

There may be a universal logic to how physicists, computers and brains tease out important features from among other irrelevant bits of data.

When in 2012 a computer learned to recognize cats in YouTube videos and just last month another correctly captioned a photo of “a group of young people playing a game of Frisbee,” artificial intelligence researchers hailed yet more triumphs in “deep learning,” the wildly successful set of algorithms loosely modeled on the way brains grow sensitive to features of the real world simply through exposure.

Using the latest deep-learning protocols, computer models consisting of networks of artificial neurons are becoming increasingly adept at image, speech and pattern recognition — core technologies in robotic personal assistants, complex data analysis and self-driving cars. But for all their progress training computers to pick out salient features from other, irrelevant bits of data, researchers have never fully understood why the algorithms or biological learning work.

Now, two physicists have shown that one form of deep learning works exactly like one of the most important and ubiquitous mathematical techniques in physics, a procedure for calculating the large-

scale behavior of physical systems such as elementary particles, fluids and the cosmos.

The new work, completed by Pankaj Mehta of Boston University and David Schwab of Northwestern University, demonstrates that a statistical technique called “renormalization,” which allows physicists to accurately describe systems without knowing the exact state of all their component parts, also enables the artificial neural networks to categorize data as, say, “a cat” regardless of its color, size or posture in a given video.

Pankaj Mehta, an assistant professor of physics at Boston University.

“They actually wrote down on paper, with exact proofs, something that people only dreamed existed,” said Ilya Nemenman, a biophysicist at Emory University. “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.”

As for our own remarkable knack for spotting a cat in the bushes, a familiar face in a crowd or

indeed any object amid the swirl of color, texture and sound that surrounds us, strong similarities between deep learning and biological learning suggest that the brain may also employ a form of renormalization to make sense of the world.

“Maybe there is some universal logic to how you can pick out relevant features from data,” said Mehta. “I would say this is a hint that maybe something like that exists.”

The finding formalizes what Schwab, Mehta and others saw as a philosophical similarity between physicists’ techniques and the learning procedure behind object or speech recognition. Renormalization is “taking a really complicated system and distilling it down to the fundamental parts,” Schwab said. “And that’s what deep neural networks are trying to do as well. And what brains are trying to do.”

Learning in Layers

A decade ago, deep learning didn’t seem to work. Computer models running the procedure often failed to recognize objects in photos or spoken words in audio recordings.

Geoffrey Hinton, a British computer scientist at the University of Toronto, and other researchers had devised the procedure to run on multilayered webs of virtual neurons that transmit signals to their neighbors by “firing” on and off. The design of these “deep” neural networks was inspired by the layered architecture of the human visual cortex — the part of the brain that transforms a flood of photons into meaningful perceptions.

When a person looks at a cat walking across a lawn, the visual cortex appears to process the scene hierarchically, with neurons in each successive layer firing in response to larger-scale, more pronounced features. At first, neurons in the retina might fire if they detect contrasts in their patch of the visual field, indicating an edge or endpoint. These signals travel to higher-layer neurons, which are sensitive to combinations of edges and other increasingly complex parts. Moving up the layers, a whisker signal might pair with another whisker signal, and those might join forces with pointy ears, ultimately triggering a top-layer neuron that corresponds to the concept of a cat.

A decade ago, Hinton was trying to replicate the process by which a developing infant’s brain becomes attuned to the relevant correlations in sensory data, learning to group whiskers with ears rather than the flowers behind. Hinton tried to train deep neural networks to do this using a simple learning rule that he and the neuroscientist Terry Sejnowski had come up with in the 1980s. When sounds or images were fed into the bottom layer of a deep neural network, the data set off a cascade of firing activity. The firing of one virtual neuron could trigger a connected neuron in an adjacent layer to fire, too, depending on the strength of the connection between them. The connections were initially assigned a random distribution of strengths, but when two neurons fired together in response to data, Hinton and Sejnowski’s algorithm dictated that their connection should strengthen, boosting the chance that the connection would continue to successfully transmit signals. Conversely, little-used connections were weakened. As more images or sounds were processed, their patterns gradually wore ruts in the network, like systems of tributaries trickling upward through the layers. In theory, the tributaries would converge on a handful of top-layer neurons, which would represent sound or object categories.

The problem was that data took too long to blaze trails all the way from the bottom network layer to the categories at the top. The algorithm wasn’t efficient enough.

Then, in 2005, Hinton and colleagues devised a new training regimen inspired by an aspect of brain development that he first learned about as a Cambridge University student in the 1960s. In

dissections of cat brains, the biologist Colin Blakemore had discovered that the visual cortex develops in stages, tweaking its connections in response to sensory data one layer at a time, starting with the retina.

To replicate the visual cortex's step-by-step development, Hinton ran the learning algorithm on his network one layer at a time, training each layer's connections before using its output — a broader-brush representation of the original data — as the input for training the layer above, and then fine-tuned the network as a whole. The learning process became dramatically more efficient. Soon, deep learning was shattering accuracy records in image and speech recognition. Entire research programs devoted to it have sprung up at Google, Facebook and Microsoft.

David Schwab, an assistant professor of physics at Northwestern University.

"In the hands of Hinton [and others], these deep neural networks became the best classifiers around," said Naftali Tishby, a computational neuroscientist and computer scientist at Hebrew University of Jerusalem. "This was very frustrating for the theoreticians in machine learning because they didn't understand why it works so well."

Deep learning worked in large part because the brain works. The analogy is far from perfect; cortical layers are more complicated than artificial ones, with their own internal networks humming away at unknown algorithms, and deep learning has branched off in directions of its own in the years since Hinton's breakthrough, employing biologically implausible algorithms for many learning tasks. But Hinton, who now splits his time between the University of Toronto and Google, considers one principle to be key to both machine and biological learning: "You first learn simple features and

then based on those you learn more complicated features, and it goes in stages."

Quarks to Tables

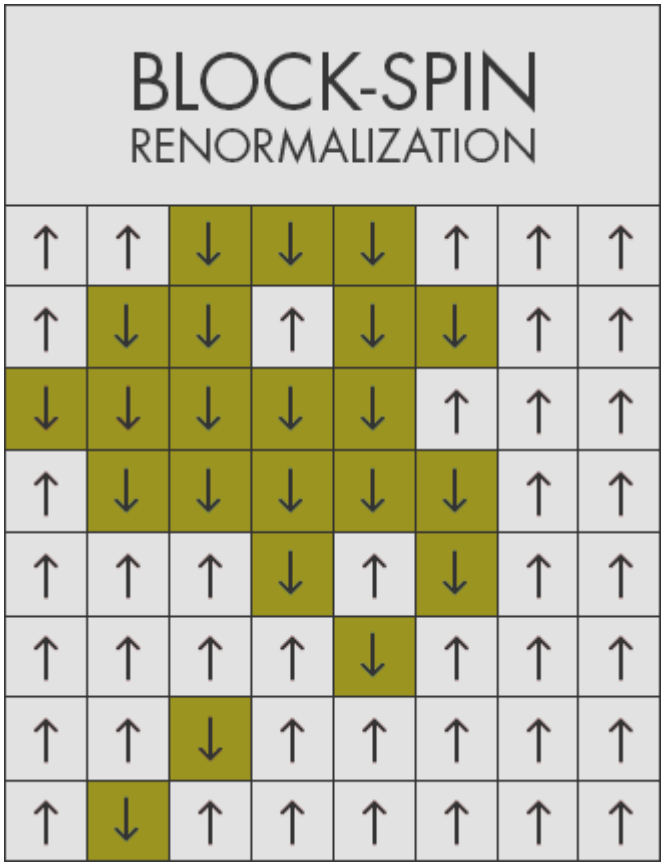
In 2010, Schwab, then a postdoctoral researcher in biophysics at Princeton University, rode the train into New York City to hear Hinton lecture about deep learning. Hinton's layer-by-layer training procedure immediately reminded him of a technique that is used all over physics and which Schwab views as "sort of the embodiment of what physics is," he said.

When he got back to Princeton, Schwab called up Mehta and asked if he thought deep learning sounded a lot like renormalization. The two had been friends and collaborators since meeting years earlier at a summer research program and frequently ran "crazy ideas" past each other. Mehta didn't find this idea particularly crazy, and the two set to work trying to figure out whether their intuition was correct. "We called each other in the middle of the night and talked all the time," Mehta said. "It was kind of our obsession."

Renormalization is a systematic way of going from a microscopic to a macroscopic picture of a physical system, latching onto the elements that affect its large-scale behavior and averaging over the rest. Fortunately for physicists, most microscopic details don't matter; describing a table doesn't require knowing the interactions between all its subatomic quarks. But a suite of sophisticated approximation schemes is required to slide up the distance scales, dilating the relevant details and blurring out irrelevant ones along the way.

Mehta and Schwab's breakthrough came over drinks at the Montreal Jazz Festival when they decided to focus on a procedure called variational or "block-spin" renormalization that the statistical physicist Leo Kadanoff invented in 1966. The block-spin method involves grouping components of a system into larger and larger blocks, each an average of the components within it. The approach works well for describing fractal-like objects, which look similar at all scales, at different levels of resolution; Kadanoff's canonical example was the two-dimensional Ising model — a lattice of "spins," or tiny magnets that point up or down. He showed that one could easily zoom out on the lattice by transforming from a description in terms of spins to one in terms of blocks of spins.

Hoping to connect the approach to the hierarchical representation of data in deep learning, Schwab and Mehta hopscotched between Kadanoff's old papers and a pair of highly cited 2006 papers by Hinton and colleagues detailing the first deep-learning protocol. Eventually, they saw how to map the mathematics of one procedure onto the other, proving that the two mechanisms for summarizing features of the world work essentially the same way.



A technique invented by Leo Kadanoff in 1966 for describing a lattice of “spins” at different levels of resolution is equivalent to a modern deep learning protocol.

To illustrate the equivalence, Schwab and Mehta trained a four-layer neural network with 20,000 examples of the Ising model lattice. From one layer to the next, the neurons spontaneously came to represent bigger and bigger blocks of spins, summarizing the data using Kadanoff’s method. “It learns from the samples that it should block-renormalize,” Mehta said. “It was astounding to us that you don’t put that in by hand, and it learns.”

A deep neural network might use a different, more flexible form of renormalization when confronted with a cat photo rather than a fractal-like lattice of magnets, but researchers conjecture that it likewise would move layer by layer from the scale of pixels to the scale of pets by teasing out and aggregating cat-relevant correlations in the data.

Summarizing the World

Researchers hope cross-fertilization between statistical physics and deep learning will yield new advances in both fields, but it is too soon to tell “what the killer app is going to be for either direction,” Schwab said.

Because deep learning tailors itself to the data at hand, researchers hope that it will prove useful for evaluating behaviors of systems that are too messy for conventional renormalization schemes, such as aggregates of cells or complex proteins. For these biological systems that lack symmetry and look nothing like a fractal, “none of the mechanical steps that we’ve developed in statistical physics work,” Nemenman said. “But we still know that there is a coarse-grained description because our own brain can operate in the real world. It wouldn’t be able to if the real world were not summarizable.”

Through deep learning, there is also the hope of a better theoretical understanding of human

cognition. Vijay Balasubramanian, a physicist and neuroscientist at the University of Pennsylvania, said he and other experts who span his two fields have long noticed the conceptual similarity between renormalization and human perception. “The development in Pankaj and David’s paper might give us the tools to make that analogy precise,” Balasubramanian said.

For example, the finding appears to support the emerging hypothesis that parts of the brain operate at a “critical point,” where every neuron influences the network as a whole. In physics, renormalization is performed mathematically at the critical point of a physical system, explained Sejnowski, a professor at the Salk Institute for Biological Studies in La Jolla, Calif. “So the only way it could be relevant to the brain is if it is at the critical point.”

There may be an even deeper message in the new work. Tishby sees it as a hint that renormalization, deep learning and biological learning fall under the umbrella of a single idea in information theory. All the techniques aim to reduce redundancy in data. Step by step, they compress information to its essence, a final representation in which no bit is correlated with any other. Cats convey their presence in many ways, for example, but deep neural networks pool the different correlations and compress them into the form of a single neuron. “What the network is doing is squeezing information,” Tishby said. “It’s a bottleneck.”

By laying bare the mathematical steps by which information is stripped down to its minimal form, he said, “this paper really opens up a door to something very exciting.”

Editor’s Note: Pankaj Mehta receives funding from the Simons Foundation as a Simons Investigator.

This article was reprinted on Wired.com.