What are the limits of deep learning?

The much-ballyhooed artificial intelligence approach boasts impressive feats but still falls short of human brainpower. Researchers are determined to figure out what's missing.

M. Mitchell Waldrop, *Science Writer*

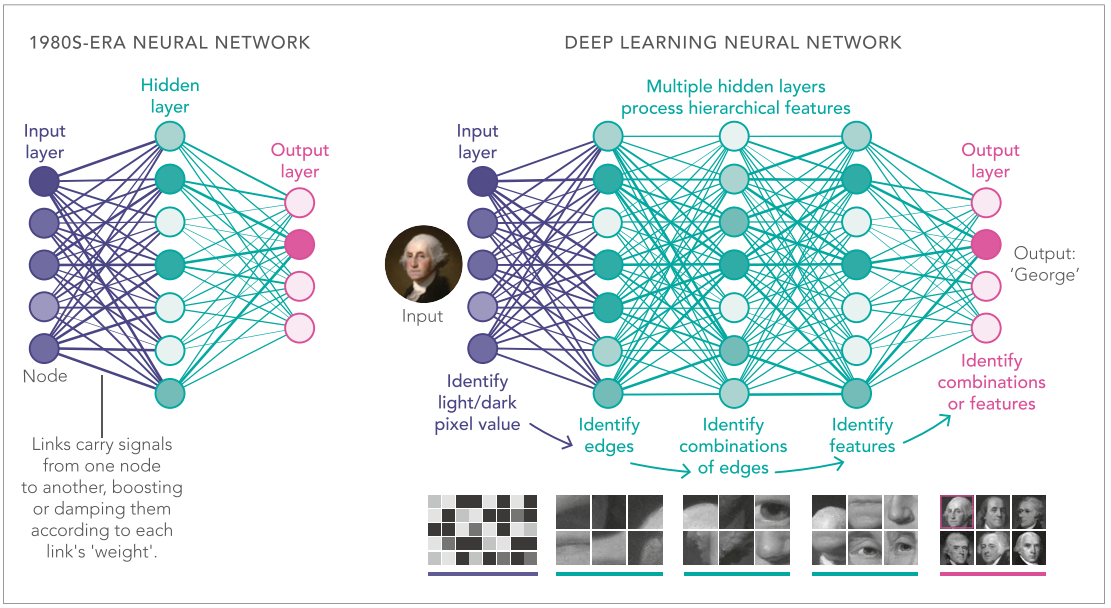
There's no mistaking the image: It's a banana—a big, ripe, bright-yellow banana. Yet the artificial intelligence (AI) identifies it as a toaster, even though it was trained with the same powerful and oft-publicized deep-learning techniques that have produced a white-hot revolution in driverless cars, speech understanding, and a multitude of other AI applications. That means the AI was shown several thousand photos of bananas, slugs, snails, and similar-looking objects, like so many flash cards, and then drilled on the answers until it had the classification down cold. And yet this advanced system was quite easily confused—all it took was a little day-glow sticker, digitally pasted in one corner of the image.

This example of what deep-learning researchers call an "adversarial attack," discovered by the Google Brain team in Mountain View, CA (1), highlights just how far AI still has to go before it remotely approaches human capabilities. "I initially thought that adversarial examples were just an annoyance," says Geoffrey Hinton, a computer scientist at the University of Toronto and one of the pioneers of deep learning. "But I now think they're probably quite profound. They tell us that we're doing something wrong."

That's a widely shared sentiment among AI practitioners, any of whom can easily rattle off a long list of deep learning's drawbacks. In addition to its vulnerability

Apparent shortcomings in deep-learning approaches have raised concerns among researchers and the general public as technologies such as driverless cars, which use deep-learning techniques to navigate, get involved in well-publicized mishaps. Image credit: Shutterstock.com/MONOPOLY919.

www.pnas.org/cgi/doi/10.1073/pnas.1821594116



"Neural network" models of AI process signals by sending them through a network of nodes analogous to neurons. Signals pass from node to node along links, analogs of the synaptic junctions between neurons. "Learning" improves the outcome by adjusting the weights that amplify or damp the signals each link carries. Nodes are typically arranged in a series of layers that are roughly analogous to different processing centers in the cortex. Today's computers can handle "deep-learning" networks with dozens of layers. Image credit: Lucy Reading-Ikkanda (artist).

to spoofing, for example, there is its gross inefficiency. "For a child to learn to recognize a cow," says Hinton, "it's not like their mother needs to say 'cow' 10,000 times"—a number that's often required for deep-learning systems. Humans generally learn new concepts from just one or two examples.

Then there's the opacity problem. Once a deep-learning system has been trained, it's not always clear how it's making its decisions. "In many contexts that's just not acceptable, even if it gets the right answer," says David Cox, a computational neuroscientist who heads the MIT-IBM Watson AI Lab in Cambridge, MA. Suppose a bank uses AI to evaluate your credit-worthiness and then denies you a loan: "In many states there are laws that say you have to explain why," he says.

And perhaps most importantly, there's the lack of common sense. Deep-learning systems may be wizards at recognizing patterns in the pixels, but they can't understand what the patterns mean, much less reason about them. "It's not clear to me that current systems would be able to see that sofas and chairs are for sitting," says Greg Wayne, an AI researcher at DeepMind, a London-based subsidiary of Google's parent company, Alphabet.

Increasingly, such frailties are raising concerns about AI among the wider public, as well—especially as driverless cars, which use similar deep-learning techniques to navigate, get involved in well-publicized mishaps and fatalities. "People have started to say, 'Maybe there is a problem'," says Gary Marcus, a cognitive scientist at New York University and one of deep learning's most vocal skeptics. Until the past year or so, he says, "there had been a feeling that deep learning was magic. Now people are realizing that it's not magic."

Still, there's no denying that deep learning is an incredibly powerful tool—one that's made it routine to deploy applications such as face and voice recognition that were all but impossible just a decade ago. "So I have a hard time imagining that deep learning will go away at this point," Cox says. "It is much more likely that we will modify it, or augment it."

Brain Wars

Today's deep-learning revolution has its roots in the "brain wars" of the 1980s when advocates of two different approaches to AI were talking right past each other.

On one side was an approach—now called "good old-fashioned AI"—that had dominated the field since the 1950s. Also known as symbolic AI, it used mathematical symbols to represent objects and the relationship between objects. Coupled with extensive knowledge bases built by humans, such systems proved to be impressively good at reasoning and reaching conclusions about domains such as medicine. But by the 1980s, it was also becoming clear that symbolic AI was impressively bad at dealing with the fluidity of symbols, concepts, and reasoning in real life.

In response to these shortcomings, rebel researchers began advocating for artificial neural networks, or connectionist AI, the precursors of today's deep-learning systems. The idea in any such system is to process signals by sending them through a network of simulated nodes: analogs of neurons in the human brain. The signals pass from node to node along connections, or links: analogs of the synaptic junctions between neurons. And learning, as in the real brain, is a matter of adjusting the "weights" that amplify or damp the signals carried by each connection.

In practice, most networks arrange the nodes as a series of layers that are roughly analogous to different

processing centers in the cortex. So a network specialized for, say, images would have a layer of input nodes that respond to individual pixels in somewhat the same way that rod and cone cells respond to light hitting the retina. Once activated, these nodes propagate their activation levels through the weighted connections to other nodes in the next level, which combine the incoming signals and are activated (or not) in turn. This continues until the signals reach an output layer of nodes, where the pattern of activation provides an answer—asserting, for example, that the input image was the number "9." And if that answer is wrong—say that the input image was a "0"—a "backpropagation" algorithm works its way back down through the layers, adjusting the weights for a better outcome the next time.

By the end of the 1980s, such neural networks had turned out to be much better than symbolic AI at dealing with noisy or ambiguous input. Yet the standoff between the two approaches still wasn't resolved—mainly because the AI systems that could fit into the computers of the time were so limited. It was impossible to know for sure what those systems were capable of.

Power Boost

That understanding began to advance only in the 2000s, with the advent of computers that were orders of magnitude more powerful and social media sites offering a tsunami of images, sounds, and other training data. Among the first to seize this opportunity was Hinton, coauthor of the backpropagation algorithm and a leader of the 1980s-era connectionist movement. By mid-decade, he and his students were training networks that were not just far bigger than before. They were considerably deeper, with the number of layers increasing from one or two to about half a dozen. (Commercial networks today often use more than 100.)

"These challenges are real, but they're not a dead end."
—Matthew Botvinick

In 2009, Hinton and two of his graduate students showed (2) that this kind of "deep learning" could recognize speech better than any other known method. In 2012, Hinton and two other students published experiments (3) showing that deep neural networks could be much better than standard vision systems at recognizing images. "We almost halved the error rates," he says. With that double whammy in speech and image recognition, the revolution in deep-learning applications took off—as did researchers' efforts to improve the technique.

One early priority was to expand the ways that deep-learning systems could be trained, says Matthew Botvinick, who in 2015 took leave from his neuroscience group at Princeton to do a year's sabbatical at DeepMind and never left. Both the speech- and image-recognition systems used what's called supervised learning, he says: "That means for every picture, there is a right answer—say, 'cat'—and if the network is wrong, you tell it what the right answer is." The network then uses the backpropagation algorithm to improve its next guess.

Supervised learning works great, says Botvinick—if you just happen to have a few hundred thousand carefully labeled training examples lying around. That's not often the case, to put it mildly. And it simply doesn't work for tasks such as playing a video game where there are no right or wrong answers—just strategies that succeed or fail.

For those situations—and indeed, for much of life in the real world—you need reinforcement learning, Botvinick explains. For example, a reinforcement learning system playing a video game learns to seek rewards (find some treasure) and avoid punishments (lose money).

The first successful implementation of reinforcement learning on a deep neural network came in 2015 when a group at DeepMind trained a network to play classic Atari 2600 arcade games (4). "The network would take in images of the screen during a game," says Botvinick, who joined the company just afterward, "and at the output end were layers that specified an action, like how to move the joystick." The network's play equaled or surpassed that of human Atari players, he says. And in 2016, DeepMind researchers used a more elaborate version of the same approach with AlphaGo (5)—a network that mastered the complex board game Go—and beat the world-champion human player.

Beyond Deep Learning

Unfortunately, neither of these milestones solved the fundamental problems of deep learning. The Atari system, for example, had to play thousands of rounds to master a game that most human players can learn in minutes. And even then, the network had no way to understand or reason about on-screen objects such as paddles. So Hinton's question remains as valid as ever: What's missing?

Maybe nothing. Maybe all that's required is more connections, more layers, and more sophisticated methods of training. After all, as Botvinick points out, it's been shown that neural networks are mathematically equivalent to a universal computer, which means there is no computation they cannot perform—at least in principle, if you can ever find the right connection weights.

But in practice, those caveats can be killers—one big reason why there is a growing feeling in the field that deep learning's shortcomings require some fundamentally new ideas.

One solution is simply to expand the scope of the training data. In an article published in May 2018 (6), for example, Botvinick's DeepMind group studied what happens when a network is trained on more than one task. They found that as long as the network has enough "recurrent" connections running backward from later layers to earlier ones—a feature that allows the network to remember what it's doing from one instant to the next—it will automatically draw on the lessons it learned from earlier tasks to learn new ones faster. This is at least an embryonic form of human-style "meta-learning," or learning to learn, which is a big part of our ability to master things quickly.

A more radical possibility is to give up trying to tackle the problem at hand by training just one big

network and instead have multiple networks work in tandem. In June 2018, the DeepMind team published an example they call the Generative Query Network architecture (7), which harnesses two different networks to learn its way around complex virtual environments with no human input. One, dubbed the representation network, essentially uses standard image-recognition learning to identify what's visible to the AI at any given instant. The generation network, meanwhile, learns to take the first network's output and produce a kind of 3D model of the entire environment—in effect, making predictions about the objects and features the AI doesn't see. For example, if a table only has three legs visible, the model will include a fourth leg with the same size, shape, and color.

These predictions, in turn, allow the system to learn quite a bit faster than with standard deep-learning methods, says Botvinick. "An agent that is trying to predict things gets feedback automatically on every time-step, since it gets to see how its predictions turned out." So it can constantly update its models to make them better. Better still, the learning is self-supervised: the researchers don't have to label anything in the environment for it to work or even provide rewards and punishments.

An even more radical approach is to quit asking the networks to learn everything from scratch for every problem. The blank-slate approach does leave the networks free to discover ways of representing objects and actions that researchers might never have thought of, as well as some totally unexpected game-playing strategies. But humans never start with a blank slate: for almost any task, they can bank on at least some prior knowledge that they've learned through experience or that was hardwired into their brains by evolution.

Infants, for example, seem to be born with many hardwired "inductive biases" that prime them to absorb certain core concepts at a prodigious rate. By the age of 2 months, they are already beginning to master the principles of intuitive physics (8), which includes the notion that objects exist, that they tend to move along continuous paths, and that when they touch they don't just pass through each other. Those same infants are also beginning to learn the basics of intuitive psychology, which includes an ability to recognize faces and a realization that the world contains agents that move and act on their own.

Having this kind of built-in inductive biasing would presumably help deep neural networks learn just as rapidly, which is why many researchers in the field are now making it a priority. Within just the past 1 or 2 years, in fact, the field has seen a lot of excitement over a potentially powerful approach known as the graph network (9). "These are deep-learning systems that have an innate bias toward representing things as objects and relations," says Botvinick.

For example, certain objects such as *paws*, *tail*, and *whiskers* might all belong to a larger object (*cat*) with the relationship *is-a-part-of*. Likewise, *Ball A* and *Block B* might have the mutual relationship *is-next-to*, the *Earth* would have the relationship *is-in-orbit-around* the *Sun*, and so on through a huge range of other examples—any of which could be represented as an abstract graph in which the nodes correspond to objects and the links to relationships.

A graph network, then, is a neural network that takes such a graph as input—as opposed to raw pixels or sound waves—then learns to reason about and predict how objects and their relationships evolve over time. (In some applications, a separate, standard image-recognition network might be used to analyze a scene and pick out the objects in the first place.)

The graph-network approach has already demonstrated rapid learning and human-level mastery of a variety of applications, including complex video games (10). If it continues to develop as researchers hope, it could ease deep learning's 10,000-cow problem by making training much faster and more efficient. And it could make the networks far less vulnerable to adversarial attacks simply because a system that represents things as objects, as opposed to patterns of pixels, isn't going to be so easily thrown off by a little noise or an extraneous sticker.

Fundamental progress isn't going to be easy or fast in any of these areas, Botvinick acknowledges. But even so, he believes that the sky's the limit. "These challenges are real," he says, "but they're not a dead end."

1 Brown TB, Mané D, Roy A, Abadi M, Gilmer J (2017) Adversarial patch. ArXiv:1712.09665 [cs.CV].
2 Mohamed A, Dahl G, Hinton G (2009) Deep belief networks for phone recognition. Available at www.cs.toronto.edu/~asamir/papers/NIPS09.pdf. Accessed December 19, 2018.
3 Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, eds Pereira F, Burges CJC, Bottou L, Weinberger KQ (Curran Associates, Inc., Red Hook, NY), Vol 25, pp 1097–1105.
4 Mnih V, et al. (2015) Human-level control through deep reinforcement learning. *Nature* 518:529–533.
5 Silver D, et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529:484–489.
6 Wang JX, et al. (2018) Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* 21:860–868.
7 Eslami SMA, et al. (2018) Neural scene representation and rendering. *Science* 360:1204–1210.
8 Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behav Brain Sci* 40:e253.
9 Battaglia PW, et al. (2018) Relational inductive biases, deep learning, and graph networks. ArXiv:1806.01261 [cs.LG].
10 Zambaldi V, et al. (2018) Relational deep reinforcement learning. ArXiv:1806.01830 [cs.LG].