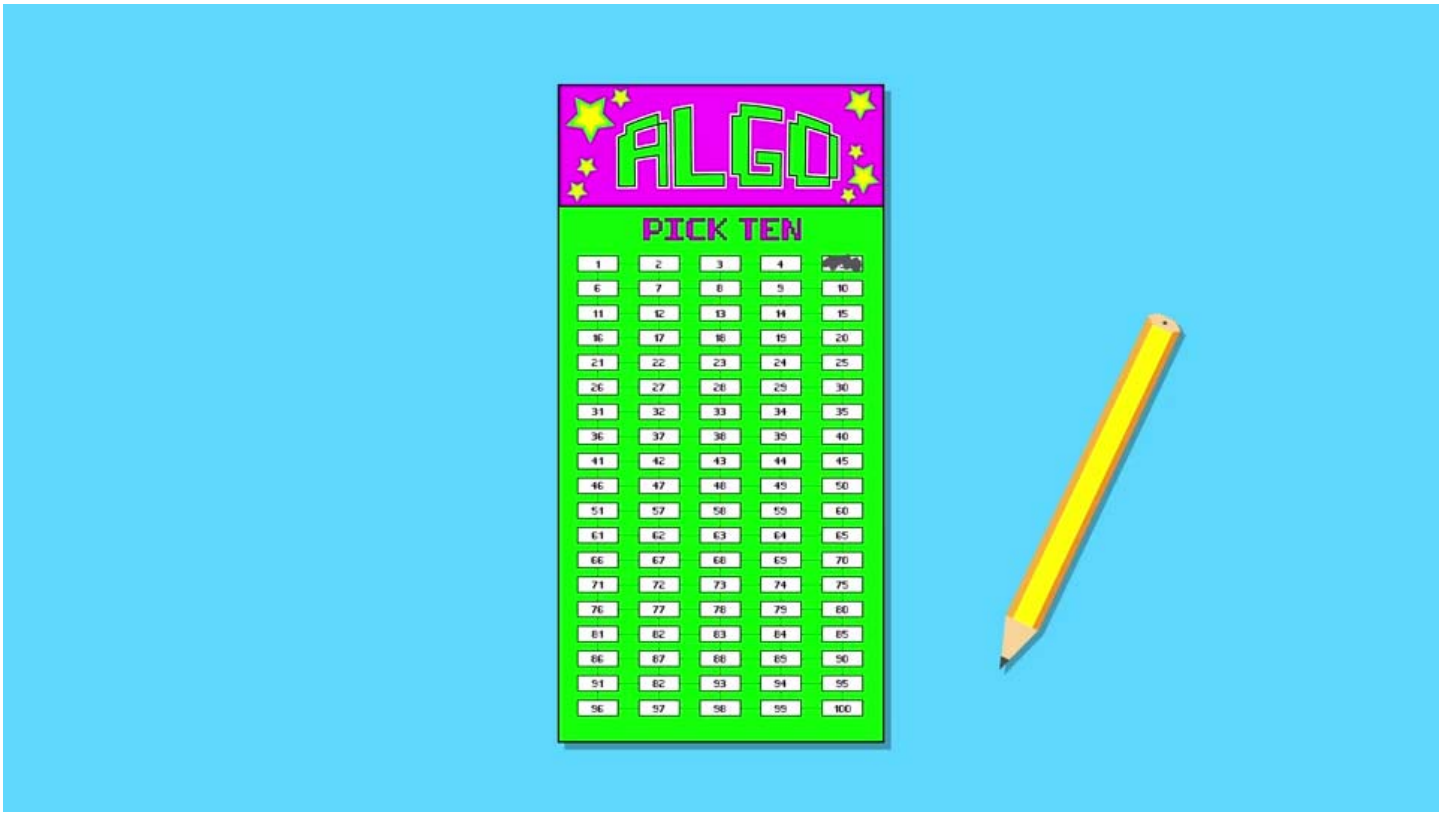
Artificial Intelligence / Machine Learning

A new way to build tiny neural networks could create powerful AI on your phone

We've been wasting our processing power to train neural networks that are ten times too big.

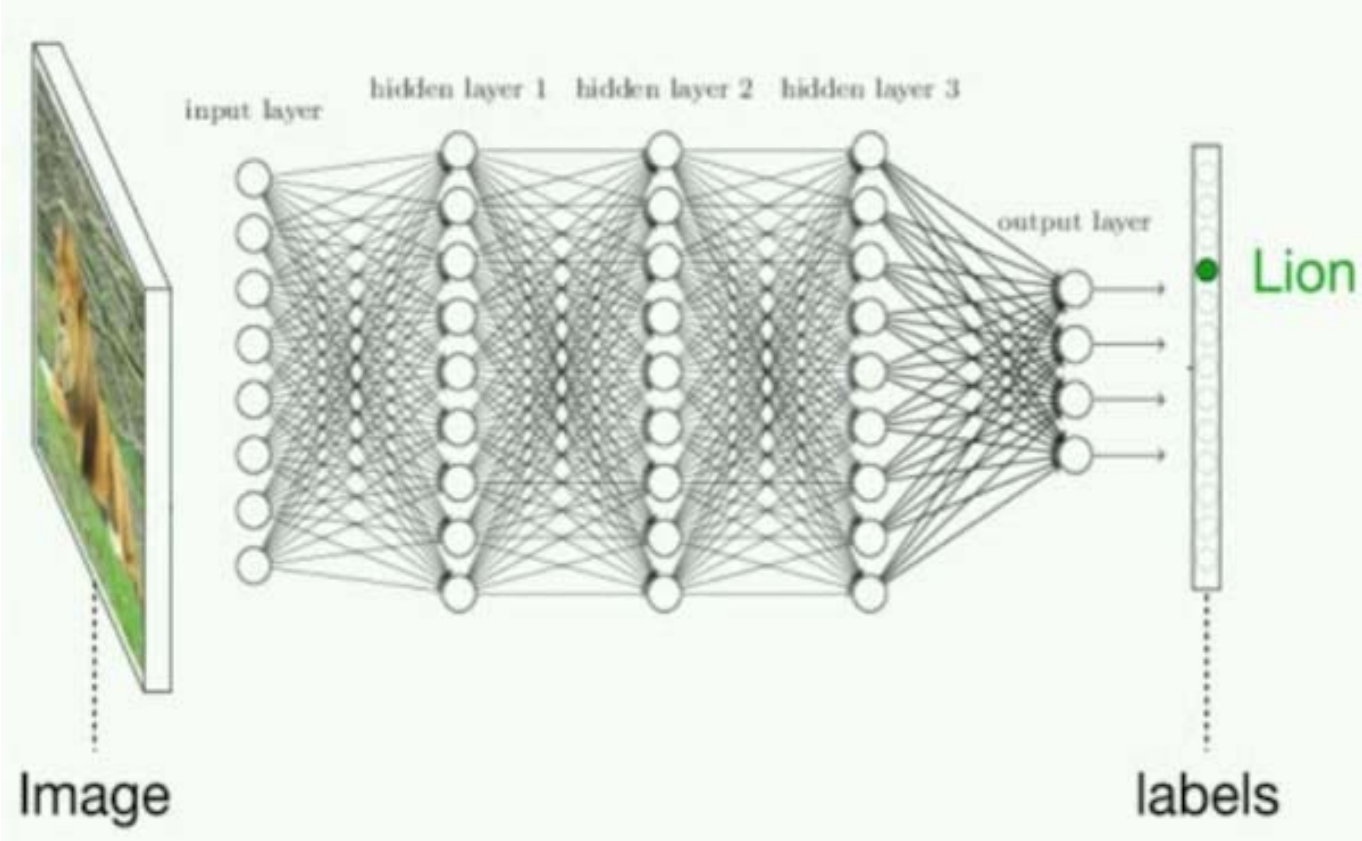
by **Karen Hao** May 10, 2019



Neural networks are the core software of deep learning. Even though they're so widespread, however, they're really poorly understood. Researchers have observed their emergent properties without actually understanding *why* they work the way they do.

Put another way, within every neural network exists a far smaller one that can be trained to achieve the same performance as its oversize parent. This isn't just exciting news for AI researchers. The finding has the potential to unlock new applications—some of which we can't yet fathom—that could improve our day-to-day lives. More on that later.

But first, let's dive into how neural networks work to understand why this is possible.



A diagram of a neural network learning to recognize a lion.
JEFF CLUNE/SCREENSHOT

How neural networks work

You may have seen neural networks depicted in diagrams like the one above: they're composed of stacked layers of simple computational nodes that are connected in order to compute patterns in data.

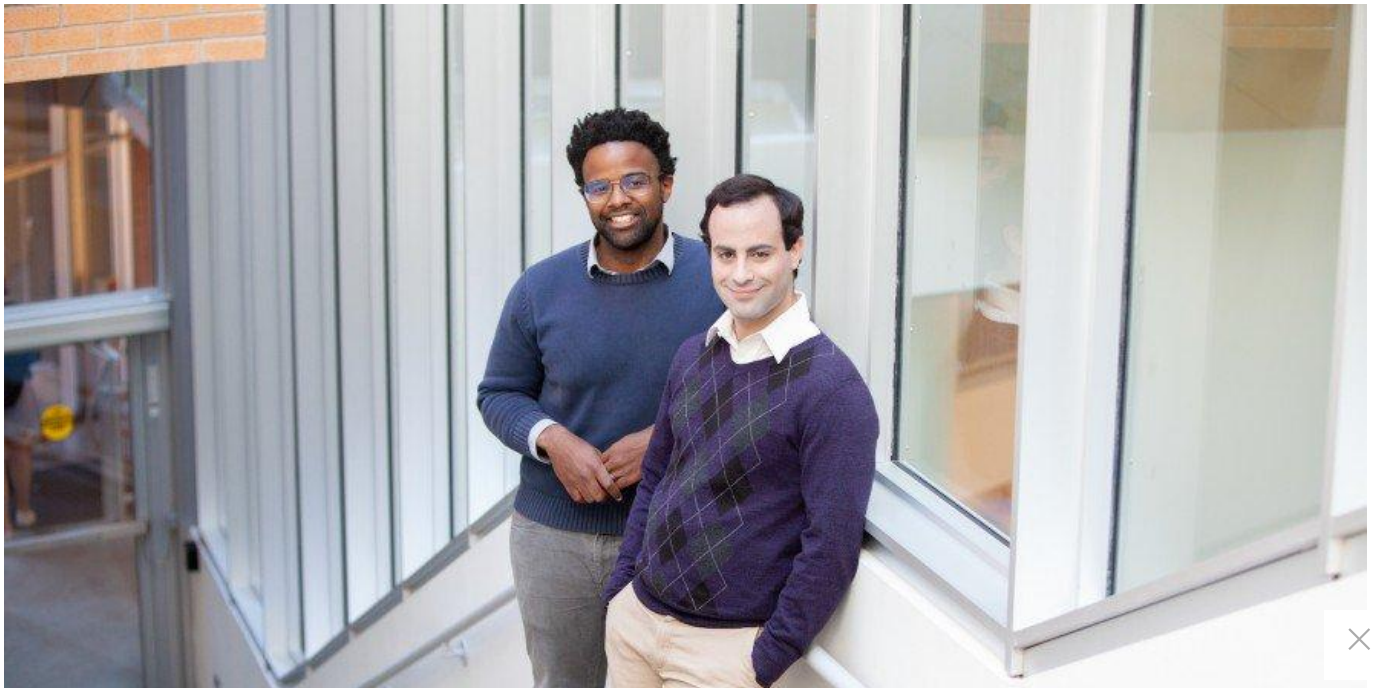
The connections are what's important. Before a neural network is trained, these connections are assigr random values between 0 and 1 that represent their intensity. (This is called the "initialization" process.)

that the randomly assigned connection strengths end up in an untrainable configuration. In other words, no matter how many animal photos you feed the neural network, it won't achieve a decent performance, and you just have to reinitialize it to a new configuration. The larger the network (the more layers and nodes it has), the less likely that is. Whereas a tiny neural network may be trainable in only one of every five initializations, a larger network may be trainable in four of every five. Again, *why* this happens had been a mystery, but that's why researchers typically use very large networks for their deep-learning tasks. They want to increase their chances of achieving a successful model.

Observation #2. The consequence is that a neural network usually starts off bigger than it needs to be. Once it's done training, typically only a fraction of its connections remain strong, while the others end up pretty weak—so weak that you can actually delete, or "prune," them without affecting the network's performance.

For many years now, researchers have exploited this second observation to shrink their networks *after* training to lower the time and computational costs involved in running them. But no one thought it was possible to shrink their networks *before* training. It was assumed that you had to start with an oversize network and the training process had to run its course in order to separate the relevant connections from the irrelevant ones.

Jonathan Frankle, the MIT PhD student who coauthored the paper, questioned that assumption. "If you need way fewer connections than what you started with," he says, "why can't we just train the smaller network without the extra connections?" Turns out you can.



The discovery hinges on the reality that the random connection strengths assigned during initialization aren't, in fact, random in their consequences: they predispose different parts of the network to fail or succeed before training even happens. Put another way, the initial configuration influences which final configuration the network will arrive at.

By focusing on this idea, the researchers found that if you prune an oversize network after training, you can actually reuse the resultant smaller network to train on new data and preserve high performance—as long as you reset each connection within this downsized network back to its initial strength.

From this finding, Frankle and his coauthor Michael Carbin, an assistant professor at MIT, propose what they call the "lottery ticket hypothesis." When you randomly initialize a neural network's connection strengths, it's almost like buying a bag of lottery tickets. Within your bag, you hope, is a winning ticket—i.e., an initial configuration that will be easy to train and result in a successful model.

This also explains why observation #1 holds true. Starting with a larger network is like buying more lottery tickets. You're not increasing the amount of power that you're throwing at your deep-learning problem; you're simply increasing the likelihood that you will have a winning configuration. Once you find the winning configuration, you should be able to reuse it again and again, rather than continue to replay the lottery.

Next steps

This raises a lot of questions. First, how do you find the winning ticket? In their paper, Frankle and Carbin took a brute-force approach of training and pruning an oversize network with one data set to extract the winning ticket for another data set. In theory, there should be much more efficient ways of finding—or even designing—a winning configuration from the start.

Second, what are the training limits of a winning configuration? Presumably, different kinds of data and different deep-learning tasks would require different configurations.

Third, what is the smallest possible neural network that you can get away with while still achieving high performance? Frankle found that through an iterative training and pruning process, he was able to consistently reduce the starting network to between 10% and 20% of its original size. But he thinks there's a chance for it to be even smaller.

Already, many research teams within the AI community have begun to conduct follow-up work. A researcher at Princeton recently teased the results of a forthcoming paper addressing the second question. A team at Uber also published a new paper on several experiments investigating the nature of the metaphorical lottery tickets. Most surprising, they found that once a winning configuration has been found, it already achieves significantly better performance than the original untrained oversize networ *before* any training whatsoever. In other words, the act of pruning a network to extract a winning

It could also change the nature of AI applications. If you can train a neural network locally on a device instead of in the cloud, you can improve the speed of the training process and the security of the data. Imagine a machine-learning-based medical device, for example, that could improve itself through use without needing to send patient data to Google's or Amazon's servers.

"We're constantly bumping up against the edge of what we can train," says Jason Yosinski, a founding member of Uber AI Labs who coauthored the follow-up Uber paper, "meaning the biggest networks you can fit on a GPU or the longest we can tolerate waiting before we get a result back." If researchers could figure out how to identify winning configurations from the get-go, it would reduce the size of neural networks by a factor of 10, even 100. The ceiling of possibility would dramatically increase, opening a new world of potential uses.

Author

Karen Hao Karen Hao is the artificial intelligence reporter for *MIT Technology Review*. In particular she covers the ethics and social impact of the technology as well as its applications for social good. She also writes the AI newsletter, the Algorithm, which thoughtfully examines the field's latest news and research. Prior to joining the publication, she was a reporter and data scientist at Quartz and an application engineer at the first startup to spin out of Google X.

03.

Bill Gates just backed a chip startup that uses light to turbocharge AI

Deepfakes June 2019

Congress is holding its first deepfakes hearing. Here's what you need to know.

With the election approaching, lawmakers are facing up to the fact they need to do something about the explosion in manipulated media.

01.
Facebook has promised to leave up a deepfake video of Mark Zuckerberg
June 2019

02.
The US military is funding an effort to catch deepfakes and other AI trickery
May 2018

03.
Inside the world of AI that forges beautiful art and terrifying deepfakes
December 2018

04.
Fake A
August

Climate Change Jun 13

Soaring temperatures will raise the risks of armed conflict

A new analysis in Nature finds that climate change has likely played a relatively small role in driving armed conflict so far. But if temperatures reach 2˚ C and 4˚ C above pre-industrial levels,...

Biotechnology Jun 13

CRISPR scientists are teaming up with a pharma giant to look for new drug clues

GlaxoSmithKline will pour $67 million into a new laboratory at the University of California to industrialize the search for drug clues using the gene-editing tool called CRISPR....

Expand

supply chains traceable.

Computing Jun 13

Telegram's boss hints that China was behind a cyberattack during Hong Kong protests

Activists have been using encrypted messaging apps like Telegram to organize demonstrations....

Expand

US elections are still far too vulnerable to attack—at every level

Cybersecurity flaws in chips are still taking too long to fix

CRISPR

A CRISPR startup is testing pig organs on monkeys

Gene-editing company eGenesis is now carrying out experiments to see if CRISPR pig organs are safe for transplant to humans.

Climate Change Jun 12

Bitcoin mining may be pumping out as much CO_2 per year as Kansas City

Read more

It's the most ambitious target set by any major economy....

Expand

CRISPR

A Russian scientist has threatened to make more CRISPR babies

01.
The search for the kryptonite that can stop CRISPR
May 2019

02.
China's CRISPR babies could face earlier death
June 2019

03.
China's CRISPR twins might have had their brains inadvertently enhanced
February 2019

04.
EXCLU are crea
Novem

Silicon Valley Jun 12

Facebook has promised to leave up a deepfake video of Mark Zuckerberg

The firm's emerging policy on deepfakes seems to be to leave them up, but flag them as fake....

Expand

Climate Change Jun 11

Cyborg seals and floating robots have solved an Antarctic ice mystery