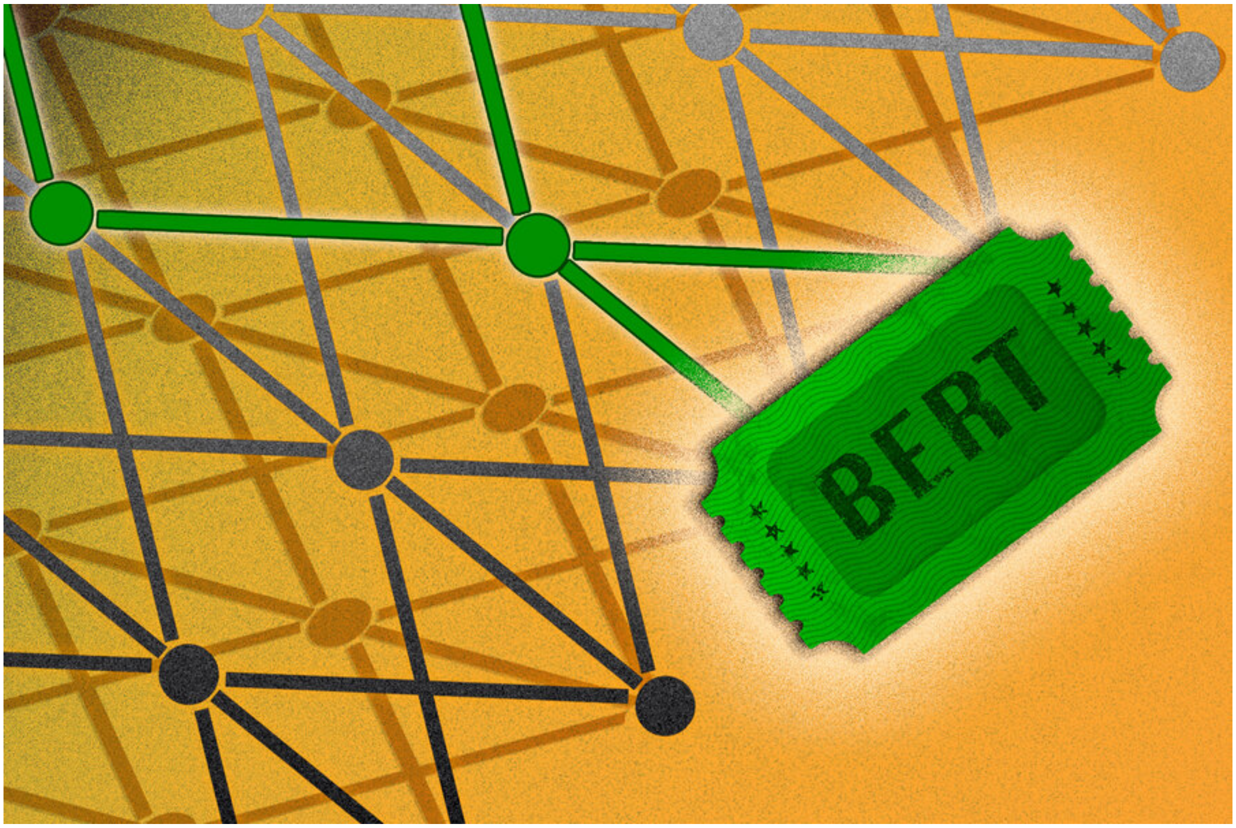
# Shrinking massive neural networks used to model language

December 1 2020, by Daniel Ackerman



Deep learning neural networks can be massive, demanding major computing power. In a test of the Lottery Ticket Hypothesis, MIT researchers have found leaner, more efficient subnetworks hidden within BERT models. Credit: Jose-Luis Olivares, MIT

You don't need a sledgehammer to crack a nut.

Jonathan Frankle is researching artificial intelligence—not noshing pistachios—but the same philosophy applies to his "lottery ticket hypothesis." It posits that, hidden within massive neural networks, leaner subnetworks can complete the same task more efficiently. The trick is finding those 'lucky' subnetworks, dubbed winning lottery tickets.

In a new paper, Frankle and colleagues discovered such subnetworks lurking within BERT, a state-of-the-art neural network approach to natural language processing (NLP). As a branch of artificial intelligence, NLP aims to decipher and analyze human language, with applications like predictive text generation or online chatbots. In computational terms, BERT is bulky, typically demanding supercomputing power unavailable to most users. Access to BERT's winning lottery ticket could level the playing field, potentially allowing more users to develop effective NLP tools on a smartphone—no sledgehammer needed.

"We're hitting the point where we're going to have to make these models leaner and more efficient," says Frankle, adding that this advance could one day "reduce barriers to entry" for NLP.

Frankle, a Ph.D. student in Michael Carbin's group at the MIT Computer Science and Artificial Intelligence Laboratory, co-authored the study, which will be presented next month at the Conference on Neural Information Processing Systems. Tianlong Chen of the University of Texas at Austin is the lead author of the paper, which included collaborators Zhangyang Wang, also of Texas A&M, as well as Shiyu Chang, Sijia Liu, and Yang Zhang, all of the MIT-IBM Watson AI Lab.

You've probably interacted with a BERT network today. It's one of the technologies that underlies Google's search engine, and it has sparked excitement among researchers since Google released BERT in 2018. BERT is a method of creating neural networks—algorithms that use layered nodes, or "neurons," to learn to perform a task through training on numerous examples. BERT is trained by repeatedly attempting to fill in words left out of a passage of writing, and its power lies in the gargantuan size of this initial training dataset. Users can then fine-tune BERT's neural network to a particular task, like building a customer-service chatbot. But wrangling BERT takes a ton of processing power.

"A standard BERT model these days—the garden variety—has 340 million parameters," says Frankle, adding that the number can reach 1 billion. Fine-tuning such a massive network can require a supercomputer. "This is just obscenely expensive. This is way beyond the computing capability of you or me."

Chen agrees. Despite BERT's burst in popularity, such models "suffer from enormous network size," he says. Luckily, "the lottery ticket hypothesis seems to be a solution."

To cut computing costs, Chen and colleagues sought to pinpoint a smaller model concealed within BERT. They experimented by iteratively pruning parameters from the full BERT network, then comparing the new subnetwork's performance to that of the original BERT model. They ran this comparison for a range of NLP tasks, from answering questions to filling the blank word in a sentence.

The researchers found successful subnetworks that were 40 to 90 percent slimmer than the initial BERT model, depending on the task. Plus, they were able to identify those winning lottery tickets before running any task-specific fine-tuning—a finding that could further minimize computing costs for NLP. In some cases, a subnetwork picked for one task could be repurposed for another, though Frankle notes this transferability wasn't universal. Still, Frankle is more than happy with the group's results.

"I was kind of shocked this even worked," he says. "It's not something that I took for granted. I was expecting a much messier result than we got."

This discovery of a winning ticket in a BERT model is "convincing," according to Ari Morcos, a scientist at Facebook AI Research. "These models are becoming increasingly widespread," says Morcos. "So it's important to understand whether the lottery ticket hypothesis holds." He adds that the finding could allow BERT-like models to run using far less computing power, "which could be very impactful given that these extremely large models are currently very costly to run."

Frankle agrees. He hopes this work can make BERT more accessible, because it bucks the trend of ever-growing NLP models. "I don't know how much bigger we can go using these supercomputer-style computations," he says. "We're going to have to reduce the barrier to entry." Identifying a lean, lottery-winning subnetwork does just that—allowing developers who lack the computing muscle of Google or Facebook to still perform cutting-edge NLP. "The hope is that this will lower the cost, that this will make it more accessible to everyone … to the little guys who just have a laptop," says Frankle. "To me that's really exciting."

**More information:** Tianlong Chen et al. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. arXiv:2007.12223 [cs.LG] arxiv.org/abs/2007.12223

*This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.*

Provided by Massachusetts Institute of Technology

Citation: Shrinking massive neural networks used to model language (2020, December 1) retrieved 1 May 2024 from https://techxplore.com/news/2020-12-massive-neural-networks-language.html