It takes a lot of energy for machines to learn: Why AI is so power-hungry

December 15 2020, by Kate Saenko

Data centers like this Google facility in Iowa use copious amounts of electricity. Credit: Chad Davis/Flickr, CC BY-SA

This month, Google forced out a prominent AI ethics researcher after she voiced frustration with the company for making her withdraw a research paper. The paper pointed out the risks of language-processing artificial intelligence, the type used in Google Search and other text analysis products.

Among the risks is the large carbon footprint of developing this kind of AI technology. By some estimates, training an AI model generates as much carbon emissions as it takes to build and drive five cars over their lifetimes.

I am a researcher who studies and develops AI models, and I am all too familiar with the skyrocketing energy and financial costs of AI research. Why have AI models become so power hungry, and how are they different from traditional data center computation?

Today's training is inefficient

Traditional data processing jobs done in data centers include video streaming, email and social media. AI is more computationally intensive because it needs to read through lots of data until it learns to understand it—that is, is trained.

This training is very inefficient compared to how people learn. Modern AI uses artificial neural networks, which are mathematical computations that mimic neurons in the human brain. The strength of connection of each neuron to its neighbor is a parameter of the network called weight. To learn how to understand language, the network starts with random weights and adjusts them until the output agrees with the correct answer.

A common way of training a language network is by feeding it lots of text from websites like Wikipedia and news outlets with some of the words masked out, and asking it to guess the masked-out words. An example is "my dog is cute," with the word "cute" masked out. Initially, the model gets them all wrong, but, after many rounds of adjustment, the connection weights start to change and pick up patterns in the data. The network eventually becomes accurate.

One recent model called Bidirectional Encoder Representations from

Transformers (BERT) used 3.3 billion words from English books and Wikipedia articles. Moreover, during training BERT read this data set not once, but 40 times. To compare, an average child learning to talk might hear 45 million words by age five, 3,000 times fewer than BERT.

Looking for the right structure

What makes language models even more costly to build is that this training process happens many times during the course of development. This is because researchers want to find the best structure for the network—how many neurons, how many connections between neurons, how fast the parameters should be changing during learning and so on. The more combinations they try, the better the chance that the network achieves a high accuracy. Human brains, in contrast, do not need to find an optimal structure—they come with a prebuilt structure that has been honed by evolution.

As companies and academics compete in the AI space, the pressure is on to improve on the state of the art. Even achieving a 1% improvement in accuracy on difficult tasks like machine translation is considered significant and leads to good publicity and better products. But to get that 1% improvement, one researcher might train the model thousands of times, each time with a different structure, until the best one is found.

Researchers at the University of Massachusetts Amherst estimated the energy cost of developing AI language models by measuring the power consumption of common hardware used during training. They found that training BERT once has the carbon footprint of a passenger flying a round trip between New York and San Francisco. However, by searching using different structures—that is, by training the algorithm multiple times on the data with slightly different numbers of neurons, connections and other parameters—the cost became the equivalent of 315 passengers, or an entire 747 jet.

Bigger and hotter

AI models are also much bigger than they need to be, and growing larger every year. A more recent language model similar to BERT, called GPT-2, has 1.5 billion weights in its network. GPT-3, which created a stir this year because of its high accuracy, has 175 billion weights.

Researchers discovered that having larger networks leads to better accuracy, even if only a tiny fraction of the network ends up being useful. Something similar happens in children's brains when neuronal connections are first added and then reduced, but the biological brain is much more energy efficient than computers.

AI models are trained on specialized hardware like graphics processor units, which draw more power than traditional CPUs. If you own a gaming laptop, it probably has one of these graphics processor units to create advanced graphics for, say, playing Minecraft RTX. You might also notice that they generate a lot more heat than regular laptops.

All of this means that developing advanced AI models is adding up to a large carbon footprint. Unless we switch to 100% renewable energy sources, AI progress may stand at odds with the goals of cutting greenhouse emissions and slowing down climate change. The financial cost of development is also becoming so high that only a few select labs can afford to do it, and they will be the ones to set the agenda for what kinds of AI models get developed.

Doing more with less

What does this mean for the future of AI research? Things may not be as bleak as they look. The cost of training might come down as more efficient training methods are invented. Similarly, while data center energy use was predicted to explode in recent years, this has not

happened due to improvements in data center efficiency, more efficient hardware and cooling.

There is also a trade-off between the cost of training the models and the cost of using them, so spending more energy at training time to come up with a smaller model might actually make using them cheaper. Because a model will be used many times in its lifetime, that can add up to large energy savings.

In my lab's research, we have been looking at ways to make AI models smaller by sharing weights, or using the same weights in multiple parts of the network. We call these shapeshifter networks because a small set of weights can be reconfigured into a larger network of any shape or structure. Other researchers have shown that weight-sharing has better performance in the same amount of training time.

Looking forward, the AI community should invest more in developing energy-efficient training schemes. Otherwise, it risks having AI become dominated by a select few who can afford to set the agenda, including what kinds of models are developed, what kinds of data are used to train them and what the models are used for.

Provided by The Conversation

Citation: It takes a lot of energy for machines to learn: Why AI is so power-hungry (2020, December 15) retrieved 21 September 2024 from https://techxplore.com/news/2020-12-lot-

energy-machines-ai-power-hungry.html