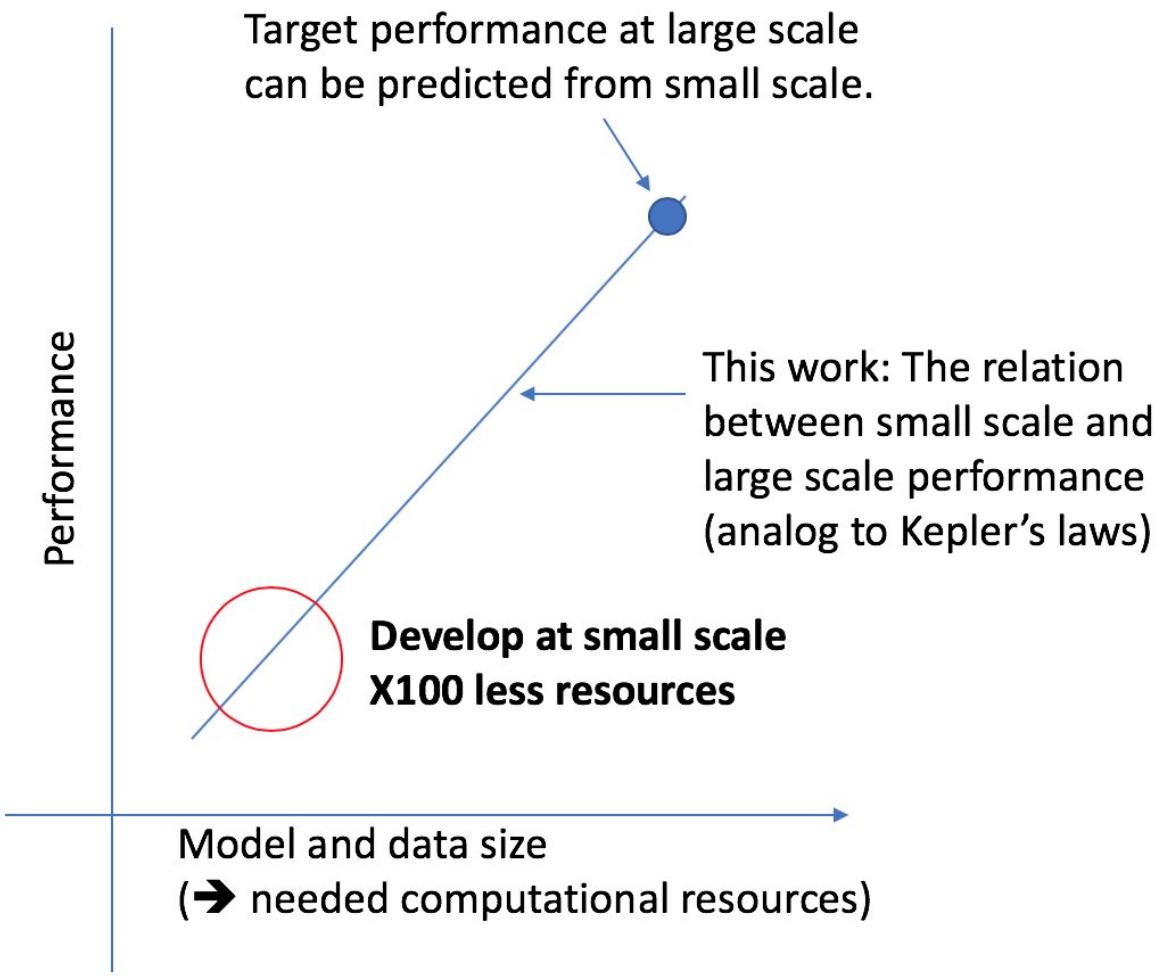
Predicting how well neural networks will scale

February 24 2020, by Adam Conner-Simons



Credit: Massachusetts Institute of Technology

For all the progress researchers have made with machine learning in helping us doing things like crunch numbers, drive cars and detect cancer, we rarely think about how energy-intensive it is to maintain the massive data centers that make such work possible. Indeed, a 2017 study predicted that, by 2025, internet-connected devices would be using 20 percent of the world's electricity.

The inefficiency of machine learning is partly a function of how such systems are created. Neural networks are typically developed by generating an initial model, tweaking a few parameters, trying it again, and then rinsing and repeating. But this approach means that significant time, energy and computing resources are spent on a project before anyone knows if it will actually work.

MIT graduate student Jonathan Rosenfeld likens it to the 17th century scientists seeking to understand gravity and the motion of planets. He says that the way we develop machine learning systems today - in the absence of such understandings - has limited predictive power and is thus very inefficient.

"There still isn't a unified way to predict how well a neural network will perform given certain factors like the shape of the model or the amount of data it's been trained on," says Rosenfeld, who recently developed a new framework on the topic with colleagues at MIT's Computer Science and Artificial Intelligence Lab (CSAIL). "We wanted to explore whether we could move machine learning forward by trying to understand the different relationships that affect the accuracy of a network."

The CSAIL team's new framework looks at a given algorithm at a smaller scale, and, based on factors like its shape, can predict how well it will perform on a larger scale. This allows a data scientist to determine if it's worth continuing to devote more resources to train the system further.

"Our approach tells us things like the amount of data needed for an architecture to deliver a specific target performance, or the most computationally efficient trade-off between data and model size," says MIT professor Nir Shavit, who co-wrote the new paper with Rosenfeld, former PhD student Yonatan Belinkov and Amir Rosenfeld of York University. "We view these findings as having far-reaching implications in the field by allowing researchers in academia and industry to better understand the relationships between the different factors that have to be weighed when developing deep learning models, and to do so with the limited computational resources available to academics."

The framework allowed researchers to accurately predict performance at the large model and data scales using fifty times less computational power.

The aspect of deep learning performance that the team focused on is the so-called "generalization error," which refers to the error generated when an algorithm is tested on real-world data. The team utilized the concept of model scaling, which involves changing the model shape in specific ways to see its effect on the error.

As a next step, the team plans to explore the underlying theories of what makes a specific algorithm's performance succeed or fail. This includes experimenting with other factors that may impact the training of deep learning models.

More information: Jonathan S. Rosenfeld, et al. A constructive prediction of the generalization error across scales. Published as a conference paper at ICLR 2020: openreview.net/pdf?id=ryenvpEKDr

Provided by Massachusetts Institute of Technology