Team develops a faster, cheaper way to train large language models

July 3 2023, by Katharine Miller

Credit: Pixabay/CC0 Public Domain

A Stanford team has developed Sophia, a new way to optimize the

pretraining of large language models that's twice as fast as current approaches.

ChatGPT and other applications that rely on large language models (LLMs) are gaining widespread use and drawing abundant media attention. But a handful of large tech companies dominate the LLM space because pretraining these models is extremely expensive, with cost estimates starting at $10 million and potentially reaching tens or hundreds of times that.

"Large language models are not very accessible to smaller organizations or academic groups," says Hong Liu, a graduate student in computer science at Stanford University.

To change that, Liu and his colleagues set out to improve on current LLM optimization methods. The result: an approach called Sophia that cuts the pretraining time in half. The details of this approach are published on the *arXiv* preprint server.

Optimizing optimization

To better optimize LLM pretraining, Liu and his colleagues, including Stanford postdoctoral fellow Zhiyuan Li, Stanford research engineer David Hall, Computer Science Assistant Professor Tengyu Ma, and Associate Professor Percy Liang, used two tricks. The first, known as curvature estimation, isn't new, but the Stanford team found a way to make it more efficient.

To understand their approach, consider a factory assembly line. To function efficiently, the factory manager needs to optimize the number of steps it takes to turn raw materials into a final product and needs to understand and appropriately staff the workload at each step along the line.

The same is true for pretraining an LLM. These models have millions or even billions of parameters that Liu likens to factory workers striving toward the same goals. One property of these parameters is their curvature, which Liu thinks of as the maximum achievable speed they reach as they progress toward the final goal of a pretrained LLM. In the factory metaphor, curvature is akin to a factory worker's workload.

If an optimization program can estimate that curvature (workload), it can make LLM pretraining more efficient. The problem is this: Estimating curvature with existing methods is remarkably difficult and expensive. "In fact, it's more expensive than doing the actual work without making curvature predictions," Liu says. That's partially why the current state-of-the-art approaches to optimizing LLM pretraining (Adam and its variants) forgo the curvature estimation step.

Still, Liu and his colleagues noticed a possible inefficiency in the prior methods that used parametric curvature estimation: Prior researchers updated their curvature estimates at every step of the optimization. The Stanford team wondered if they could make the process more efficient by decreasing the number of updates.

To test that idea, the Stanford team designed Sophia to estimate parameters' curvature only about every 10 steps. "That turned out to be a huge win," Liu says.

The team's second optimization trick, called clipping, addresses a related issue: The problem of inaccurate curvature estimation. "If the estimation is wrong, it's like giving people with hard jobs even more work to do. It makes things worse than if there were no estimation at all."

Clipping prevents that by setting a threshold, or a maximum curvature estimation. "In our factory metaphor, it's like setting a workload limitation for all employees," Liu says. Another metaphor often applied

to optimization is a landscape of hills and valleys where the goal is to end up in the lowest valley. Without clipping, Liu says, it is possible to land at a saddle between two mountains. "In optimization, that's not where you want to be," he says.

Testing Sophia and scaling up

Liu and his colleagues used Sophia to pretrain a relatively small LLM using the same model size and configuration that were used to create OpenAI's GPT-2.

Sophia's combination of curvature estimation and clipping allowed the LLM pretraining optimization to smoothly proceed to the lowest valley in half the number of steps and half the time required by Adam.

"Sophia's adaptivity sets it apart from Adam," Liu says. "It's harder for Adam to handle parameters with heterogeneous curvatures because it can't predict them in advance."

It's also the first time in nine years that anyone has shown any substantial improvement over Adam on language model pretraining, Liu says. "This could mean a huge reduction in the cost of training real-world large models." And as models scale, Sophia's advantages should only increase, he says.

Next, Liu and his colleagues hope to develop a larger LLM using Sophia. He's also hoping to see Sophia applied to other areas of machine learning such as computer vision models or multi-modal models. "It would take some time and resources to move Sophia to a new domain, but because it is open source, the community could certainly do it."

More information: Hong Liu et al, Sophia: A Scalable Stochastic Second-order Optimizer for Language Model Pre-training, *arXiv* (2023).

DOI: 10.48550/arxiv.2305.14342

Provided by Stanford University