IBM Newsroom

Kailash Gopalakrishnan

IBM Distinguished Research Staff Member, Senior Manager of Accelerator Architectures and Machine Learning

High-performance computing usually relies on precise mathematics. So, you might think that an approximate calculation wouldn't work on something as complex as machine learning.

But Kailash Gopalakrishnan and his colleagues at IBM Research found that being close is as good as being exact when training deep learning models—the backbone of AI systems today. Working off a technique known as "approximate computing," they found a way for computer chips to efficiently process machine learning data with increased computing power that enable users—from the cloud to the edge—to do more with less.

"It's the most exciting project I've worked on at IBM," Kailash says. It's also the type of work he had dreamed of doing when he was a student earning his bachelor's degree in electrical engineering from the Indian Institute of Technology, in Bombay, and his doctorate in the field from Stanford University.

"I always knew I wanted to be an engineer," says Kailash, who lives in New York City and works at the IBM Thomas J. Watson Research Center in Yorktown Heights. "I was fascinated by mathematics and physics, and that journey led me to AI."

Since he joined IBM in 2004, his research has relied on the disciplines of engineering, physics and mathematics. Over the last seven years he's looked for new ways to make AI more computationally efficient and consumable. "We're putting together complex hardware and software built from the ground up for artificial intelligence," he says.

The Power of Imprecision

Companies are racing to develop innovative methods to build hardware that can meet the soaring computational needs driven by advancements in AI. Kailash and his IBM Research colleagues recognized that AI applications are resilient to imprecise computations, which, in turn, could be extremely efficient in hardware. The team's efforts led to the creation of multiple generations of highly cited IBM AI accelerator chips that are highly efficient at scaled precision arithmetic.

"When you bring approximate computing techniques such as low-precision arithmetic to AI problems,'' he explains, "you have the opportunity to improve the performance of building and deploying AI models by several orders of magnitude."

Approximate computing can accomplish deep learning training and inference tasks at not just 16 or 32 bits, but even down to just a few bits without losing fidelity. Their research isn't "a one-trick pony," Kailash adds. The research team has already demonstrated there is a roadmap for precision scaling ideas that can continue to provide growth in AI data throughput while fully preserving the accuracy of large complex models.

Kailash and his colleagues' accomplishments have stirred a great deal of interest inside and outside IBM. Last year, in partnership with the state of New York, IBM Research launched the AI Hardware Center, a global research hub focused on enabling next-generation chips and systems that support the tremendous processing power and unprecedented speed that AI requires to realize its full potential.

The goal, Kailash explains, is to broaden the ecosystem and applications by marrying AI software and hardware. As the field progresses, organizations around the world will be able to derive myriad benefits from fast and affordable machine learning.

As Kailash accepts the IBM Fellow title, he hopes to be known for “solving really hard problems that helped lay the foundation on which future AI innovations are built.”

→ ***Kailash Gopalakrishnan on Linkedin***

→ ***Meet the next IBM Fellow,Shalini Kapoor.***

https://newsroom.ibm.com/kailash