PURDUE INNOVATES | Office of Technology Commercialization

Weight Streaming

New methods reduce GPU memory needs for advanced deep learning by optimizing data storage and utilizing higher CPU memory.

Researchers at Purdue University have developed two novel methods for reducing GPU memory requirements for reverse automatic differentiation. Neural networks are increasing in layers (getting deeper) and nodes (getting wider), but GPUs have limited memory, limiting the run time of differentiable programs. Even with increased capacity, sometimes multiple GPUs are required for processing. The methods developed by Purdue researchers improve the running of extremely deep neural networks and extremely long-running differential programs. One of the methods, termed divide and conquer checkpointing, reduces the memory requirement for storing the intermediate values and results. This is particularly useful for reverse automatic differentiation, which requires saving intermediate results of the forward sweep to perform the reverse sweep. Another method, termed tensor streaming, performs just-in-time migration of data back and forth between the CPU and GPU. This utilizes the higher memory of CPUs compared to GPUs; the highest-performing CPUs (8 TB) have 100 times more memory in a single node than the highest-performing GPUs (80 GB).

Technology Validation: Evaluation of the researchers' system in various real-world examples showed highly efficient use of CPU and GPU resources.

Advantages

- Reduces GPU memory requirements

Applications

- Training large deep-learning models

TRL: 3

Intellectual Property:

Provisional-Gov. Funding, 2021-09-10, United States | PCT-Gov. Funding, 2021-12-23, WO | NATL-Patent, 2021-12-23, Europe | NATL-Patent, 2024-03-

Technology ID
2021-SISK-69295

Category
Artificial Intelligence & Machine Learning/AI Model Optimization & Acceleration Tools

Authors
Hamad Ahmed
Jeffrey M Siskind

Further information
Dipak Narula
dnarula@prf.org

View online

06, United States

Keywords: GPU memory reduction, reverse automatic differentiation, neural networks, deep learning models, divide and conquer checkpointing, tensor streaming, differentiable programs, CPU GPU migration, deep neural networks, intermediate value storage, Automatic Differentiation, Computer Technology, deep learning, Differentiable Programming, GPU, Machine Learning