FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning

Wednesday, 5 October 2022 14:15 (15 minutes)

Applications of machine learning (ML) are growing by the day for many unique and challenging scientific applications. However, a crucial challenge facing these applications is their need for ultra low-latency and on-detector ML capabilities. Given the slowdown in Moore's law and Dennard scaling, coupled with the rapid advances in scientific instrumentation that is resulting in growing data rates, there is a need for ultra-fast ML at the extreme edge. Fast ML at the edge is essential for reducing and filtering scientific data in real-time to accelerate science experimentation and enable more profound insights. To accelerate real-time scientific edge ML hardware and software solutions, we need well-constrained benchmark tasks with enough specifications to be generically applicable and accessible. These benchmarks can guide the design of future edge ML hardware for scientific applications capable of meeting the nanosecond and microsecond level latency requirements. To this end, we present an initial set of scientific ML benchmarks, covering a variety of ML and embedded system techniques.

Primary authors: HAWKS, Benjamin (Fermi National Accelerator Lab); HERWIG, Christian (Fermi National Accelerator Lab. (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); MUHIZI, Jules (Fermilab/Harvard University); TRAN, Nhan (Fermi National Accelerator Lab. (US)); PRAKASH, Shvetank (Harvard University); JANAPA REDDI, Vijay (Harvard University)

Presenter: MUHIZI, Jules (Fermilab/Harvard University)

Session Classification: Contributed Talks