***Snowmass Letter of Interest*, Topic Areas: CompF03, IF04, IF07**

Fast Machine Learning

Maria Acosta Flechas[4], Markus Atkinson[1], Giuseppe Di Guglielmo[8], Javier Duarte[2], Farah Fahim[4], Philip Harris[3], Christian Herwig[4], Burt Holzman[4], Ryan Kastner[2], Mia Liu[5], Chang-Seong Moon[7], Mark Neubauer[1], Kevin Pedro[4], Andres Quintero-Parra[4], Dylan Rankin[3], Ryan Rivera[4], Nhan Tran[4,6], Michael Wang[4], Tingjun Yang[4], Joshua Agar[9], and Eliu A. Huerta[1]

[1]University of Illinois Urbana Champaign, Champaign, IL 61820, USA
[2]University of California San Diego, La Jolla, CA 92093, USA
[3]Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Fermi National Accelerator Laboratory, Batavia, IL 60510, USA
[5]Purdue University, West Lafayette,IN 47907, USA
[6]Northwestern University, Evanston, IL 60208, USA
[7]Kyungpook National University, Daegu, 41566, Korea
[8]Columbia University, New York, NY 10027, USA
[9]Lehigh University, Bethlehem, PA, 18015, USA

ABSTRACT

Machine learning (ML) methods have been actively explored as novel solutions for simulation, reconstruction and analysis in fundamental particle physics research. Accelerated ML inference offers numerous opportunities to leverage these advanced algorithms to enhance the instrumentation and computing capabilities of particle physics experiments, and provides a computing ethos for diverse fields where low-latency inference is required. Low-latency applications realized on FPGAs, GPUs, and ASICs can improve trigger performance and minimize data loss. In bringing fast ML as close as possible to sensor data, we can drastically improve the data rate which can be streamed off the detector. In addition, the use of heterogeneous computing resources to accelerate offline reconstruction and simulation may serve as an elegant solution to computing challenges faced by many particle physics experiments. Furthermore, fast ML may improve experimental operation and accelerator and detector control without drastic overhauls in the design. In this letter of interest, we summarize explorations by the community, and highlight prospects in future developments and applications.

1 Introduction

Fundamental particle physics experiments are taking data with unprecedented rate and volume, constantly challenging our online and offline data processing capabilities. At the CERN Large Hadron Collider (LHC), protons are accelerated and collided at a rate of approximately 40 MHz at each interaction point, making it impossible to read out and save all collision events detected by the ATLAS and CMS experiments, which produce $\mathcal{O}(100)$ Tb/s of data. During the online data collection, the event processing and filtering begins on-detector in application-specific integrated circuits (ASICs) and off-detector in field-programmable gate arrays (FPGAs) with communication and readout via optical fibers. After this first stage of filtering, known as the Level-1 trigger (L1T), the data is passed to a computing farm to be further analyzed and filtered in the high-level trigger (HLT), where accelerated computing is critical to meet the throughput and latency requirements for data acquisition. The offline processing of the massive data sets saved, such as the HL-LHC data expected in the future (> 1 EB), also present high-throughput computing challenges. The big data challenge is not limited to the LHC and its high-luminosity upgrade (HL-LHC): neutrino experiments, such as the Deep Underground Neutrino Experiment (DUNE), and cosmology experiments like the Vera C. Rubin Observatory, will produce datasets comparable in size to the HL-LHC dataset. Additionally, there is increased demand for real-time processing of data to identify supernovae, neutron star mergers, and other multi-messenger astronomical signatures across multiple experimental domains, including neutrino physics, astrophysics, and gravitational wave astronomy. This effort synergies well with Department of Energies AI for Science report which highlights the need for Edge computing broadly across scientific thrusts.[1] Furthermore, as the collision environments get busier with more simultaneous interactions (as expected at the HL-LHC and other future colliders) and astrophysics experiments gain in sensitivity, the need for more sophisticated and powerful algorithms, including machine learning (ML), will increase. Such challenges are faced by many particle physics experiments across the field and are discussed in Ref. [2] in greater detail. The history, motivation, and reinvigorated interest in ML for high energy physics (HEP) is broad. Existing reviews [3–7] have laid out the impact of ML on detector and accelerator controls and operation, data simulation, reconstruction, and analysis. In this letter of interest, we discuss explorations of accelerated ML inference as solutions to our online and offline data processing challenges.

2 Ultrafast Machine Learning Inference with `hls4ml`

Recent ML developments exploit the capability of deep neural networks and have gained in popularity inside and outside of the HEP community. To realize these methods on specialized hardware such as FPGAs and ASICs, the challenges are multifaceted. The large number of computations make it a non-trivial task to program on energy-efficient and fast computing hardware, which often requires specialized engineering skills to program. Thus, programming these specialized resources may be costly and time-consuming. The `hls4ml` project aims to provide a versatile tool for users in the science community to program such devices and evaluate the performance, latency, and resource usage of their algorithms. The development cycle of algorithm design, firmware conversion, emulation, and evaluation is streamlined with `hls4ml`. In Ref. [8], a jet tagging network ($> 5,000$ connections) programmed on an FPGA was demonstrated to meet the ultralow-latency (< 100 ns) requirement of the LHC L1T. In this work, network compression, including iterative pruning and quantization, were employed to achieve the desired performance and resource usage. The trade-off between network performance and FPGA resource usage was also investigated.

Recently, this effort was expanded to support more algorithms, such as boosted decision trees [9] and graph neural networks [10]. Quantization-aware training with QKeras [11] was also investigated. Binary and ternary networks were also studied and showed a large reduction in DSP usage, which is often the limiting resource on the FPGA [12]. The `hls4ml` tool has also been applied extensively for tasks in the HL-LHC upgrade of the CMS L1T system, including an autoencoder for anomaly detection, and DNNs for muon energy regression and identification, tau lepton identification, and vector boson fusion event classification [13]. The codesign concept has also been expanded to design an ASIC that encodes sensor data as a reduced representation which compresses the data volume read off the chip.

3 Heterogeneous Compute Accelerated Machine Learning Inference

Fundamental particle physics has pushed the bounds of computing for decades. Our current computing model primarily utilizes on-premises computing resources in data centers sited at national laboratories and university campuses. As single-threaded performance of CPUs has plateaued and datasets continue to grow by orders of magnitude, more efficient and specialized types of computing architectures need to be explored. As a result, heterogeneous computing resources—mixed-architecture computing systems that can offer large gains in performance—are becoming increasingly popular to meet data processing demands. We explore various applications for accelerated ML inference at several experiments. The future computing challenges of CMS and ATLAS, both in online streaming HLT applications and offline raw processing, are significantly greater than the resources currently being deployed. This presents an opportunity for transformative changes to the computing model and technology.

In the Fast ML group, we have explored utilizing heterogeneous computing resources, often optimized for ML inference, *as a service* to process our event-based data. This approach allows non-disruptive integration of external computing resources in a multi-threaded computing model with complex job scheduling. We demonstrated a 4–100 times speedup with Microsoft brainwave FPGA-coprocessor service compared to a CPU implementation, with a top tagging example based ResNet-50 [14]. This work has been extended to GPUs-as-services and other FPGA services such as Amazon Web Services (AWS) Elastic Compute Cloud (EC2) F1 FPGA instances [15]. We also expanded the number of applications including HCAL energy regression at CMS HLT and a DeepCalo model in ATLAS.

Accelerated ML inference that can scale to processing of large data volumes will also be important for offline reconstruction and selection of neutrino interactions. DUNE will conduct a rich program with 30 PB of raw data anticipated to be collected per year. It will be a challenge to efficiently analyze that data-set without transformations in computing models and technology that can handle data retrieval, transport, paralleled processing, and storage in a cohesive manner. In Ref. [16], we demonstrate the speedup of neutrino interaction reconstruction accelerated with GPU co-processors as-a-service.

Gravitational-wave astrophysics as enabled by the LIGO and Virgo detectors is another area where discovery is challenged by the computational requirements in order to regress noise artifacts (utilizing thousands of auxiliary detector channels) and carry out the astrophysical searches in real-time. The first observation via gravitational waves as well as multi-wavelength electromagnetic observations of a binary neutron star system in galaxy NGC 4993 has established the field of multi-messenger astrophysics with gravitational waves. Building on our work in heterogeneous, real-time computing, we have started an effort to incorporate machine learning and hardware accelerators in gravitational wave data analysis problems as well[17]. We expect this to impact low-latency discovery of gravitational-wave sources and their mutli-messenger follow-up[18].

4 Outlook and Opportunities

In this letter of interest, we outlined a number of the present and upcoming TDAQ and computing challenges for a several fundamental physics experiments. This includes both ultrafast online filtering and real-time streaming computing applications as well as large data-set processing and simulation performed offline. A few explorations in exploiting accelerated ML as solutions to these challenges were presented above. In addition to expanding the capabilities of the tools we support, the objectives of the studies are also being extended beyond latency and resource usages: energy awareness fast ML inference is also being explored in the group, expanding the application domain of the fast ML inferences. The fast ML efforts will build foundations for future intelligent detectors capable of self-adjusting in an ever changing experimental environment or fully autonomous operation control systems. Furthermore, solving problems under stringent system specifications also offers unique opportunities for cross cutting research in understanding AI method information content as well as their interpretability.

Given these challenges, we believe there is the potential for great impact in fundamental physics in deploying new technologies. Across the various experimental applications, there are some generic considerations when approaching these opportunities.

- **Generalizable hardware implementation of models**: While ML is a powerful tool, the exploration of such techniques is constantly evolving and the types of network architectures (and thus specialized computations) will also necessarily change. Physicists are often creating custom network architectures which presents challenges for generalizability of hardware implementations.

- **Elastic, non-disruptive**: A considerable amount of the current infrastructure and software has been developed around the (event-based) computing model for fundamental physics. Being able to incorporate and elastically extend that computing model, for example through services and container orchestration, is an important consideration in how best to deploy emerging computing hardware

- **Global infrastructure**: For large international experiments, the computing resources are often globally distributed across many countries (certainly true for the LHC experiments). This will challenge not only raw computing power but also networking bandwidth, data storage, and datacenter orchestration and operation.

- **Continuous learning**: Most applications in physics are currently feed-forward “fire and forget” inference. For future applications of autonomous operations, controls, and calibration, system level design requires a feedback loop such that the algorithms can learn from the data to adapt to changing conditions and continual optimization

- **Community Building**: The Fast ML group will build a community of developers with diverse use-cases and expertise. This will allow creative cross-fertilization of new ideas and provide a collective for collaborative open source developments. As one tangible example members of the team are involved in directing future mid-scale infrastructure investments by NSF in electron microscopy and gravitational waves.[19, 18]

The intersection of **ML inference, intelligent sensors, and accelerated compute is a very exciting opportunity** with rapid developments in both academia and industry, and fundamental physics is an area which presents interesting applications to **explore emerging technologies**.

References

1. "AI for Science Report". https://www.anl.gov/ai-for-science-report. Accessed: 2020-8-31.
2. M. Atkinson et al., "Opportunities for Accelerated Machine Learning Inference in Fundamental Physics", 02, 2020. doi:10.5281/zenodo.4001022.
3. K. Albertsson et al., "Machine Learning in High Energy Physics Community White Paper", in *Proceedings, 18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2017)*, volume 1085, p. 022008. 2018. arXiv:1807.02876. doi:10.1088/1742-6596/1085/2/022008.
4. HEP Community, "A living review of machine learning for particle physics", 2020. https://iml-wg.github.io/HEPML-LivingReview/.
5. A. Radovic et al., "Machine learning at the energy and intensity frontiers of particle physics", *Nature* **560** (2018) 41, doi:10.1038/s41586-018-0361-2.
6. D. Guest, K. Cranmer, and D. Whiteson, "Deep learning and its application to LHC physics", *Ann. Rev. Nucl. Part. Sci.* **68** (2018) 161, doi:10.1146/annurev-nucl-101917-021019, arXiv:1806.11484.
7. D. Bourilkov, "Machine and deep learning applications in particle physics", *Int. J. Mod. Phys. A* **34** (2020) 1930019, doi:10.1142/S0217751X19300199, arXiv:1912.08245.
8. J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", *J. Instrum.* **13** (2018) P07027, doi:10.1088/1748-0221/13/07/P07027, arXiv:1804.06913.
9. S. Summers et al., "Fast inference of boosted decision trees in FPGAs for particle physics", *J. Instrum.* **15** (2020) P05026, doi:10.1088/1748-0221/15/05/P05026, arXiv:2002.02534.
10. Y. Iiyama et al., "Distance-weighted graph neural networks on FPGAs for real-time particle reconstruction in high energy physics", (2020). arXiv:2008.03601. Submitted to *Front. Big Data*.
11. C. N. Coelho et al., "Ultra low-latency, low-area inference accelerators using heterogeneous deep quantization with QKeras and `hls4ml`", (2020). arXiv:2006.10159.
12. G. Di Guglielmo et al., "Compressing deep neural networks on FPGAs to binary and ternary precision with `hls4ml`", *Mach. Learn.: Sci. Technol.* (2020) doi:10.1088/2632-2153/aba042, arXiv:2003.06308.
13. CMS Collaboration, "The Phase-2 upgrade of the CMS Level-1 trigger", CMS Technical Design Report CERN-LHCC-2020-004. CMS-TDR-021, CERN, 2020.
14. J. Duarte et al., "FPGA-accelerated machine learning inference as a service for particle physics computing", *Comput. Softw. Big Sci.* **3** (2019) 13, doi:10.1007/s41781-019-0027-2, arXiv:1904.08986.
15. J. Krupa et al., "GPU coprocessors as a service for deep learning inference in high energy physics", (7, 2020). arXiv:2007.10359. Submitted to *Mach. Learn: Sci. Technol.*
16. M. Wang et al., "GPU-accelerated machine learning inference as a service for computing in neutrino experiments", (2020). In Preparation.
17. R. Ormiston et al., "Noise reduction in gravitational-wave data via deep learning", *Physical Review Research* **2** (Jul, 2020) doi:10.1103/physrevresearch.2.033066.
18. "NSF Award Search: Award#1931469 and #1931561 - Collaborative Research: Frameworks: Machine learning and FPGA computing for real-time applications in big-data physics experiments". https://www.nsf.gov/awardsearch/showAward?AWD_ID=1931469&HistoricalAwards=false. Accessed: 2020-8-31.
19. "NSF Award Search: Award#2038140 - MsRI-EW: Enabling Transformative Advances in Materials Engineering through Development of Novel Approaches to Electron Microscopy". https://www.nsf.gov/awardsearch/showAward?AWD_ID=2038140&HistoricalAwards=false. Accessed: 2020-8-31.