BNL High Energy Physics AI/ML Vision

November 2021

Machine learning techniques have been successfully used in HEP and other sciences for over 30 years. The deep learning revolution since a decade ago and consequent rapid growth in AI/ML techniques, accompanying software tools and leveraged hardware is creating tremendous potential for enhancement in AI/ML scientific computing tools with multiple benefits to HEP.

Our vision for AI/ML is two-fold. We seek to exploit the development of established and emerging AI/ML techniques to solve experiment and theory driven computing challenges, and we seek to develop AI/ML techniques assuring that the application of AI/ML to our problems is sound, with quantifiable uncertainties. We see AI/ML as a tool and approach to apply after exhausting well understood "conventional" ways to extract information from data. With this strategy we use accrued domain knowledge to wisely guide the application of AI/ML, to produce results that can be understood and trusted, together with their uncertainties.

AI/ML is being widely investigated and used today across BNL HEP programs, as an important tool capable of improving our physics results. Our activities and near-term plans are a mix of exploratory R&D, specific applications, tools development, and planning/developing towards future applications, with all research programs having identified AI/ML activities in separate FWPs totaling about $1 million. In the Energy Frontier AI/ML techniques are being applied to improve Higgs identification and the precision of the Higgs mass, cross sections, and couplings measurements; to identify longitudinally polarized W bosons both for current ATLAS measurements and for future runs/colliders; to further improve the energy calibration of jets, the location of electrons, and the momentum of muons; and in the trigger to reconstruct tracks from charged particles the new baseline for HL-LHC. Our ATLAS program is also developing highly scalable AI/ML workflow services to provide large-scale processing that can empower sophisticated AI/ML applications within ATLAS and beyond. Our DUNE effort is engaged in a software and computing program that is applying existing AI/ML techniques to ProtoDUNE data and exploring novel techniques to develop calibration strategies ultimately for application to DUNE data. Further Intensity Frontier effort is generating realistically simulated AI/ML training data sets of LArTPC waveforms and reconstructed 3D images; using AI/ML to improve Wire-Cell’s neutrino vertex identification efficiency in MicroBooNE; using AI/ML to improve particle flow reconstruction in Wire-Cell; and developing and improving infrastructure software for high-performance computing

facilities (HPC/CPU/GPU) in support of AI/ML training and inference. These activities have strong synergies with a BNL-funded AI/ML LDRD project 'LS4GAN' described below. In Cosmic Frontier AI/ML approaches are being applied to astronomical optical image analysis (deblending, deconvolution, denoising), and self-learning RFI rejection in digital electronics. Additional AI/ML activities include HEP theory developing algorithms for lattice gauge theory computations; detector R&D developing real time AI/ML techniques to improve detector and trigger response; and an early career award program (Tricoli) developing AI/ML algorithms to select ATLAS events in which the Higgs is produced through vector boson fusion and decays into two W bosons.

An example of work delivering both the near and long term is ATLAS simulation. At the HL-LHC ATLAS aims to use fast simulation for 90% of its simulation statistics, saving time relative to full Geant4 but making fast simulation a substantial processing consumer in itself. The fast simulation processing chain is projected to use 21% of ATLAS processing by 2030. Using ML to improve both the speed and accuracy of the fast simulation is making good progress. BNL-led work on speeding up the FastCaloGAN calorimeter simulation, the first ML based simulation code to enter ATLAS production, has lately improved training speed by 2-3x and evaluation speed by 5x. ML based simulation can leverage large heterogeneous HPCs, which are increasingly designed and optimized for good AI/ML performance. With one FastCaloGAN training cycle taking ~1 GPU-month of processing time, ATLAS is already presenting significant ML processing needs that can benefit from large scale resources.

In an example of applying AI/ML to the understanding of simulation uncertainties, a new R&D program at BNL ('LS4GAN') makes use of Generative Adversarial Networks (GANs) to train a model that translates simulated HEP data to real data. A given simulated event will then differ from its translated version in ways that can reveal inaccuracies, artifacts, resolutions, and biases in the simulation. Furthermore, with the difference between simulated events and their translated form representing the modelling deficiency with respect to actual data, the simulated events and their translated counterparts can be propagated through the full offline software chain with the difference between final outputs representing the overall detector simulation uncertainty. The data volumes and neural networks involved are large, requiring large scale processing for training and inference that will benefit from HPCs and the BNL-developed services to use them.

BNL HEP AI/ML activities are strengthened and complemented by activities elsewhere in the laboratory; an example is a new R&D collaboration between ATLAS and EIC physicists to apply AI/ML intelligence at the detector level. With LHC data rates at the PB/s scale, annual processed data volumes already exceeding an exabyte, and storage the largest cost component of LHC computing (and inevitably large for future facilities as

well), intelligent filtering at the experiment's DAQ system can yield large economies in downstream costs without degrading the physics. BNL is applying its expertise to develop this approach, using modern deep learning techniques on powerful commodity FPGAs to achieve ~100x reduction in data storage requirements.

In our AI/ML vision, our focus in the near term (1-2 years) is on building up community, scientific applications, and the tools to empower them. AI/ML coordination within HEP and at the laboratory level facilitates communication, collaboration and 'training up' the staff.

Activities and deliverables (representative subset), current/near term:

- Communication and collaboration channels established at department and laboratory level
    - Department level AI/ML meeting series and Mattermost channel
    - Laboratory level AI/ML Working Group developing journal club series, seminars, workshops, tutorials
- AI/ML based study of simulation systematics applied to the first target application of ProtoDUNE LAr TPC simulation
- Develop AI/ML methods as growing contributor to faster and more accurate ATLAS simulation
- BNL-developed distributed AI/ML services facilitating the use of large scale heterogeneous HPCs for simulation production

Our midterm (3-5 year) focus will be on the maturation of current and emerging R&D, and the laboratory infrastructure integration. R&D topics being pursued on this timescale via LDRDs, DOE's SciDAC program and other mechanisms include the mentioned simulation systematics study (LDRD) and AI/ML intelligence at the detector (LDRD), real-time particle tracking with deep learning on FPGAs (LDRD), image classification in the HEP cosmic frontier (SciDAC), an Intelligent Data Delivery Service supporting large scale AI/ML workflows (US ATLAS HL-LHC computing R&D), and AI/ML based operational intelligence in workflow management (ATLAS computing R&D). Most of these studies benefit from strong engagement with various organizations at BNL and beyond.

Activities and deliverables (representative subset), 3–5 year term:

- The Laboratory will deploy a common computing infrastructure supporting AI/ML on this timescale and laboratory-wide committee to accomplish this is formed

    - We will integrate HEP activities with this infrastructure, with our large-scale AI/ML workflow services serving our science applications running on BNL infrastructure, DOE HPCs and other facilities
- Extend application of AI/ML in ATLAS simulation beyond FastCaloGAN to the FastChain workflow targeted as the basis for 90% of HL-LHC simulation as part of the BNL-led US ATLAS HL-LHC computing R&D program
- Extend simulation uncertainties studies beyond ProtoDUNE to other experiments
- On this timescale the DAQ system design collaboration with EIC aims to deliver an AI/ML based tool chain implemented in firmware and software to optimize data taking in response to detector conditions

Our longer term (5-10 year) vision focuses on consolidating BNL HEP as a leader in using AI/ML to increase the quality and productivity of scientific output across HEP program, and in making BNL HEP a leader in the curation and use of high energy physics datasets for AI/ML and other studies. BNL is and will be a steward of crucial data-intensive experimental datasets (ATLAS including HL-LHC, Belle II, RHIC, EIC) that is a resource for scientists employing AI/ML techniques for ongoing scientific discovery, facilitated by processing services that build on our distributed AI/ML activities.