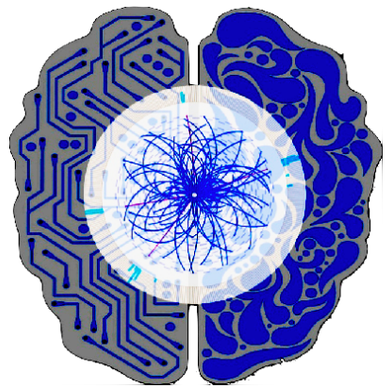

AI Hardware Overlap

CompF3 (ML)

Dylan Rankin [MIT]

April 7th, 2022

Introduction

- CompF3 group focused on machine learning
- Wide range of white paper topics
  - Some strong overlap with CompF4
- Will try to discuss possible areas of overlap, relevant points

CompF3 WPs (1)

- G. Kasieczka, B. Nachman, D. Shih, et al. “The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics”, arXiv:2101.08320

- S.V. Chekanov, W. Hopkins, “Event-based anomaly detection for new physics searches at the LHC using machine learning”, arXiv:2111.12119 [hep-ph]

- **Denis Boyda, Salvatore Calì, Sam Foreman, Lena Funcke, et al. “Applications of Machine Learning to Lattice Quantum Field Theory”. arXiv:2202.05838**

- Alexander Scheinker, Spencer Gessner. ”Adaptive Machine Learning for Time-Varying Systems: Towards 6D Phase Space Diagnostics of Short Intense Charged Particle Beams“, arXiv: 2203.04391 [physics.acc-ph]

- Brett Viren, Jin Huang, Yi Huang, Meifeng Lin, Yihui Ren, Kazuhiro Terao, Dmitrii Torbunov, Haiwang Yu. ”Solving Simulation Systematics in and with AI/ML”, arXiv:2203.06112

- Alexander Bogatskiy, Sanmay Ganguly, Thomas Kipf, Risi Kondor, David W. Miller, et al. ”Symmetry Group Equivariant Architectures for Physics”, arXiv:2203.06153

- Daniel Diaz, Javier Duarte, Sanmay Ganguly, Raghav Kansal, Samadrita Mukherjee, Brian Sheldon, Si Xie. ”Improving Di-Higgs Sensitivity at Future Colliders in Hadronic Final States with Machine Learning“, arXiv:2203.07353

CompF3 WPs (2)

- **Cora Dvorkin, Siddharth Mishra-Sharma, Brian Nord, V. Ashley Villar, Camille Avestruz, et al. ”Machine Learning and Cosmology”, arXiv:2203.08056**

- **Rainer Bartoldus, Catrin Bernius, David W. Miller. ”Innovations in trigger and data acquisition systems for next-generation physics facilities“, arXiv:2203.07620 (also in IF04)**

- **Andreas Adelmann, Walter Hopkins, Evangelos Kourlitis, Michael Kagan, Gregor Kasieczka, et al. ”New directions for surrogate models and differentiable programming for High Energy Physics detector simulation”, arXiv:2203.08806**

- N. Akchurin, J. Damgov, S. Dugad, P. G C, S. Grönroos, K. Lamichhane, J. Martinez, T. Quast, S. Undleeb, A. Whitbeck. ”Deep learning applications for quality control in particle detector construction”, arXiv:2203.08969

- **Savannah Thais, Paolo Calafiura, Grigorios Chachamis, Gage DeZoort, Javier Duarte, et al. ”Graph Neural Networks in Particle Physics: Implementations, Innovations, and Challenges”, arXiv:2203.12852**

- **Philip Harris, Erik Katsavounidis, William Patrick McCormack, Dylan Rankin, Yongbin Feng, et al. ”Physics Community Needs, Tools, and Resources for Machine Learning”, arXiv:2203.16255**

Lattice Quantum Field Theory

- *Applications of Machine Learning to Lattice Quantum Field Theory, arXiv: 2202.05838*

- What is good for ML is good for LQFT (mostly)

- GPUs see significant use already
    - Mention of IPUs & TPUs

- LQFT requires more tightly interconnected nodes than typical ML applications
    - Large models needing fast communications are becoming increasingly common

- Typical calculations require double precision
    - Mixed-precision algorithms can be employed to allow efficient use of hardware

Cosmology

- *Machine Learning and Cosmology, arXiv:2203.08056*
- GPUs for training (certainly not unique)
  - Expect increase in GPU/HPC usage
- TPUs and FPGAs for follow-up, pre-processing for fast detection
  - Could be part of an array at data processing facility
  - Could be located at the observational facility to bypass data transfer times

HEP Detector Simulation

- *New directions for surrogate models and differentiable programming for High Energy Physics detector simulation, arXiv:2203.08806*

- Take advantage of common ML tool support for GPUs

Graph Neural Networks

- *Graph Neural Networks in Particle Physics: Implementations, Innovations, and Challenges, arXiv:2203.12852*

- GNNs (for physics) not as well supported as other architectures (although improving)

    - Optimized libraries for sparse data on GPU may not be suitable

    - Physics application may need 4(+) dimensions

- Acceleration of GNN training and inference on coprocessors is promising direction for future R&D

    - GPUs, FPGAs, IPUs, Intel Habana Goya & Gaudi cards, TPUs, Cerebras

- Certain type of GNN (GarNet) implemented in hls4ml (for FPGAs), necessary for low-latency usage

Needs, Resources, Tools (1)

- *Physics Community Needs, Tools, and Resources for Machine Learning, arXiv:2203.16255*

- Covers types of hardware, software for usage, resources for access

  - Examples from colliders, neutrinos, astrophysics

  - GPUs expected for large offline use-cases (ex. event reconstruction, noise estimation, signal detection/extraction)

  - FPGAs already in use for low-latency (ex. LHC/DUNE trigger, astro alerts)

  - ASICs an option for high-rad, low-power environments (ex. collider front-ends)

  - Exploration of IPUs, TPUs, optical processors should be given attention

Needs, Resources, Tools (2)

- Focus on the tools to use this hardware
- HLS has enabled a lot of prototyping for ML
- Electronic design automation (EDA) tools necessary for synthesis/integration
    - ASIC tools/licenses can be very expensive (joint agreements?)
- Open source tools are really crucial for collaboration, adapting to specific needs

Needs, Resources, Tools (3)

- Access to AI hardware is also critical
- SONIC/Nvidia Triton/as-a-service tools use hardware efficiently w/o significant effort/expertise
    - Hardware not required to be local
- Cloud computing is scalable
    - Costly for sustained usage, costs may decrease in the future
- HPC centers large resource
    - Not optimally designed for all use cases (ex. CPU:GPU ratio 2:1 not appropriate for aaS model)