Collider AI topics

Tim, Alex, Monica, Benedikt, Nikos

Disclaimer:

Talks in last workshop, white paper and examples still a good summary

– Software, Hardware and ML-Ops

– Fast-ML/FPGA Discussion

– White paper

Try to update and give direction of travel since last year

Not a comprehensive overview, just fuel for discussion …

Direction of travel in (some) offline use cases

Reconstruction & object ID

– Last year mainly **GNN** based examples: jet (flavour) tagging, track reconstruction, …

– Exploration of **transformer** architectures in the meantime, for jet tagging (e.g. ParT), track reconstruction, calorimeter clustering …

– **Bayesian NN** with uncertainty estimates from ensembles (e.g. ATLAS Calo)

– Various challenges including **training** of large models and **scaling** of inference times etc. – in the end a slow march towards foundation models? (e.g. dataset)

Direction of travel in (some) offline use cases

Analysis level use cases

– Many (most?) analyses use **BDT**s (especially LHCb) or **DNN**s for high-level analysis needs e.g. signal/background event classification

– Limited number of high-level variables, **quick**, **easy** to train …

– Emerging analysis techniques potential to increase complexity, e.g. Simulation Based Inference, normalising flows …

– Analysis use cases catch up in **complexity** and **computational** needs?

Direction of travel in (some) offline use cases

Other use cases

– Anomaly detection (events/jets…) increasingly important/popular topic

– Autoencoders to find unusual events through compression/reconstruction loss

– Fast (detector) simulations using generative AI e.g FlashSim

– Add your favourite here …

Direction of travel in (some) offline use cases

Tools and communication etc.

– Collaborations mainly use industry tools – interfaced to experiment software directly or via intermediate tools e.g. ONNX

– Some custom internal frameworks (e.g. Salt) – no huge push to align internally/across experiments?

– Getting more serious about ML Ops (see later comments and workshop)

– Experimental fora for discussion: conferences, workshops, LPCC IML …

– Models trained/developed typically with CERN/institute resources – UK?

Direction of travel in real time use cases

FastML (typically FPGA/GPU, low-latency use cases for triggering)

– Similar evolution towards attention/transformers/AD as offline in last year

– Different (larger?) challenges in real time world e.g. latency, chip resource ..

– Strong progress on tools:

  – High Granularity Quantisation (HGQ) becoming widely used

  – Distributed Arithmetic (DA) becoming widely used

  – Investigation of custom AI engines in FPGAs

  – Development of CI/CD pipelines and ML ops – critical in trigger world

Direction of travel in real time use cases

Communication

– UK well aligned with **Next Gen Trigger** project at CERN (ATLAS > CMS)

– UK part of wider **FastML community** – https://fastmachinelearning.org

– Community tools e.g. hls4ml, conifer, HGQ, da4ml … some UK contribution?

– Non-profit to be set up in US to support tools long term (currently NGT, A3D3…)

– Support from industry (AMD, Altera, Nvidia, Google…)

– More intra UK communication? Opportunities to build on strong UK expertise?

Direction of travel in real time use cases

Opportunities

- EPIGRAPHY (Edge deeP learnInG foR pArticle PHYsics)
  - UK lead: Benedikt Maier (Imperial) & Caterina Doglioni (deputy, Manchester)
- MLQC4FC (Machine Learning and Quantum Computing for Future Colliders)
  - UK leads: Valentina Vecchio (Manchester) & Simon Williams (Durham)
- ML4FE (Machine Learning for Front End)
  - White paper planned (synergy with UK ASIC expertise/DRD?)

The end …

What else has changed/evolved? What did I forget?

Have we managed to make progress on the recommendations from last year?

What key things should we aim to do in the coming months/year?

Recommendations

1. It is essential for the UK to adopt a strategic approach to AI for HEP, including sustained funding, skills development, talent retention, knowledge transfer, industry collaboration, and participation in global AI-initiatives.
2. Provision of dedicated, stable and long-term funding to ensure the UK is at the forefront of harnessing AI, whilst increasing the socioeconomic impact of the field.
3. Establish a regular forum for discussing common barriers, challenges, and opportunities in AI across the HEP field. An in-person meeting twice a year, circulating around different venues, was decided as an appropriate forum. Working groups focused on specific areas can meet more frequently as appropriate.
4. Increase engagement of UK-based researchers in AI strategy initiatives beyond-HEP, ensuring the UK's interests are represented and aligned with the outcomes.
5. Centrally maintain and advertise details of AI hardware and development environments available to HEP researchers.
6. Share expertise, document best practices and investigate shared frameworks/tools to aid knowledge exchange within the UK.
7. Develop a coherent approach to AI training in HEP in the UK, for researchers of all career stages, including:
    a. More stable and focussed delivery of STFC's data-science summer school.
    b. Build upon successful HEP AI training programmes, such as the CDTs, to ensure successful elements are more widely adopted and benefits widely felt.
    c. Initiate focused AI software carpentry-style workshops.
    d. Encourage more expert-exchange placement, with industry and within HEP.
8. Enhance AI knowledge exchange in the UK by mapping expertise, creating common forums/software, enabling expert practitioners to work across experiments, and building strong cross-experimental cohorts.
9. To better exploit their deep foundational expertise, build stronger links between HEP researchers and STFC's national facilities such as Hartree Centre and the SciML.
10. Where most essential, strengthen links to foundational AI expertise beyond HEP, by identifying and building links with motivated foundational researchers.
11. Ensure HEP is sufficiently attractive to retain/attract AI experts, in terms of job progression, security and making an attractive offering. Explore innovative offerings, such as joint industry and research roles.
12. Identify applications or projects (individuals, collaborative or consortia) that could be quickly initiated for future wider funding calls.
13. Exploit more efficient industry tools to maximally leverage developments and position for expertise exchange.
14. Maximise the wider socio-economic impact of HEP, enabling the field to tap into innovation funds and to build a stronger narrative for long-term investment from both government and industry.
15. Build on the UK HEP community's expertise in high-speed electronics and data processing algorithms to further develop Fast-ML activities, ensuring strong interconnections through a UK forum for groups of experts or interested parties.
16. Continue to develop leading-edge solutions in μs-scale AI, growing industry engagement, and making outcomes available to other research areas and beyond.
17. Commit to further contributions to development, support and training for open-source toolsets to enable fast AI, primarily on FPGA, for instance hls4ml and conifer.