ML4EP
2024 Plan of Work

ML4EP Meeting - 16/12/2024

Lorenzo Moneta

ML4EP

- A new initiative for common ML activities within SFT started this year
- Built to federate, consolidate and coordinate existing ML activities within SFT projects:
  - ML for fast simulation
    - developments of models for fast simulation of calorimeter showers
  - ML software in ROOT
    - interfaces for using external ML software within ROOT (e.g. Batch generator)
    - C++ inference of ML models (SOFIE)
- ML activities of NGT WP1
  - hls4ml (ML for FPGA)
  - Model compression and training
  - Interfaces for ML inference

Fast Simulation: Future Plans

- Continue exploring new generative ML techniques (rapidly evolving)
- Further developments of CaloDiT
  - optimise architecture, lighter attention mechanism, hyperparameter optimisation
  - model inference optimisation (distillation, flow-matching,...)
- Continue the collaboration with experiments
  - ATLAS:
    - test CaloDiT on ATLAS on both EM and hydronic showers.
    - improve energy modelling with potential flow model on top of CaloDiT
  - CMS: test CaloDiT in HGCal
  - LHCb: support the implementation of CaloDiT
  - FCCee: Ensure Par04 models are working with FCCeeCLD and FCCeeALLEGRO
  - Open Data Detector: dataset generation and support FastCaloSim demonstrator
  - Community effort: next edition of Calo Challenge
- Exploit point/cloud representation for showers (summer student??)
- NTNU project and collaboration with Linz?

Future Plans for ML Software (SOFIE)

- SOFIE
  - Implement new operators according to needs
    - already have a large number of operators
    - can parse ParticleNet, GNN1 from ATLAS and (almost) distilled diffusion model
    - expressed interest in using SOFIE also from non-CERN experiments (Belle-II, ePIC)
  - Memory and CPU optimisations
    - better usage of memory in computational graph
  - GPU porting (e.g. using ALPAKA?)
    - interest from CMS to have optimal interface to ML inference
  - interoperability with HLS4ML?

Future Plans for ML Software

- Maintain benchmark of different ML inference solutions
- Model repository in cvmfs
- Promote the batch generator as a convenient interface for training
  - integrate into the currently developed training framework
    - b-hive from CMS and Salt/FTAG from ATLAS
    - integration with ml.cern.ch (based on kubeflow)
- Support ML workflows for Simulation-Based Inference
  - Integration of ML with statistical tools (RooFit)

NGT Plans - 1.7

- Interfaces for ML inference
  - not clear milestones defined in NGT proposal in first 2 years
- Will start discussions with experiments (CMS and LHCb) on possible activities
  - Investigate possible solutions
    - SOFIE (with porting to GPU with ALPAKA)
    - Interfaces to TensorRT and ROCm for NVidia and AMD GPUs
    - Explore new possibilities (e.g. AITemplate)
  - Develop benchmarks for the different solutions using HEP models

NGT Plans - 1.2

- Milestones from NGT proposal

Time	Description	Deliverable/Milestone
6 m	Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	Integration in hls4ml on multiple backends
12 m	- Deployment of transformers on FPGAs - Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	- Integration in hls4ml on multiple backends - Journal publication on Knowledge Distillation on Transformer use case
18 m	Support for generic Graph Neural Networks	- Improved code-generation infrastructure to support general graphs on multiple hls4ml backends - Journal publication on Graph NN fast inference
24 m	- Support for generic Transformer network - Mid-point hls4ml release	- Journal publication describing novel hls4ml functionalities and example applications - Tutorial describing new hls4ml functionalities

NGT Plans - 1.3

- Milestones from NGT proposal

Time	Description	Deliverable/Milestone
6 m	Baseline development: large-scale training and optimization workflow on at least one end-to-end training library (Pytorch/Tensorflow)	Integration of the developed algorithms on the NNLO library (large-scale training package for CERN custom training workflow on HPC infrastructure)
12 m	Support of optimal workflows for hardware-aware pruning techniques with resource estimation.	- Demonstrator of network training and architecture scan for a concrete benchmark use case from WP2 or WP3 - NNLO tutorial showcasing novel functionalities - Journal publication
18 m	Support for Knowledge Distillation at training	integration of the developed compression workflows in the NNLO library
24 m	- AutoML-like flow towards automatic optimization of quantization and pruning at training time - Application of hardware-aware training on real-life use cases from WP2 and WP3	- Mid-point NNLO software release - Journal publication - NNLO tutorial showcasing novel functionalities

POW 2025

- We could follow ROOT reporting
- Make initial Google spreadsheet with the different items
  - add the priority
  - length of tasks (e.g. S, M, L) ?
  - assign a person responsible for the task
- Monitor during the year looking if a task is
  - DONE, PARTIALLY DONE or NOT DONE

Backup Slides

Machine Learning

LM, JR

See Lorenzo's talk Vision for a new ML/AI activity !

Priority 1:

- Put RBatchGenerator in production
- Consolidate RBDT
- Support of integration of SOFIE in experiments Fast Simulation pipelines
- Add support in SOFIE for NVidia GPUs in CUDA
- Continue to add support for the ONNX operators requested by experiments

Priority 2:

- Make HLS4ML interoperable with SOFIE
- Streamline ROOT's inference interface, making it able to use models for Python ML frameworks (e.g. Keras/TF) directly

We want to support experiments inference (C++) for cases that are difficult to implement or require heavy dependencies.

We don't want to compete with existing industry tools for training.

from A. Ribon presentation

Fast Simulation

The ML-related work items will be integrated into the new ML activity

- Develop transformer-based ML models
  - Establish the best single-geometry diffusion model
  - Work on inference optimisation
  - Extend to different geometries and test adaptation capabilities, measure savings on training time
- Experiment-specific work (in collaboration with members of the experiments)
  - LHCb
    - Find the best working model for hadronic showers (possibly a transformer-based model)
  - ATLAS
    - New Fellow (Peter Mckeown) will continue the work of D. Salamani on ML for ATLAS, implementing a data structure that allows to test VAE and transformer-based models
    - Co-supervise work of J. Beirer on FastCaloSimV2-based classical shower simulation
  - CMS
    - Implement data production sample with structure that allows to test transformer-based models on HGCal
- Others
  - Speed-up simulation of oriented crystals detector
  - Community efforts : CaloChallenge and Open Data Detector