Vision for a new AI/ML Activity in SFT

SFT PoW for 2024

Lorenzo Moneta *15/1/2024*

New ML Initiative in SFT

- A new initiative for common ML activities within SFT
- Build on existing ML activities within SFT projects:
  - ML for fast simulation
    - developments of models (VAE, transformers) for fast simulation of calorimeters
  - ML software in ROOT
    - interfaces for using external ML software within ROOT (e.g. Batch generator)
    - C++ inference of ML models (SOFIE)
  - *See the detailed plans for these activities in the ROOT and Simulation presentations*

Project Goals for SFT

- Promote collaboration between different SFT projects on AI/ML topics
  - sharing expertise and knowledge
  - using common solutions within SFT for Machine Learning
  - propose and investigate new ideas for using AI/ML in SFT projects
- Develop and maintain common ML software solutions required for experiments. Some possible examples:
  - ML models for fast simulation of calorimeters
  - inference for heterogeneous architectures (CPU, GPU, FPGA)
  - interfaces and integration with HEP software
  - software pipeline for training models (improve SWAN integration)

Project Vision

- A place for sharing common AI/ML expertise within SFT and its stakeholders
- Foster collaboration with experiments (where major ML developments are happening) and also with IT (innovation group and OpenLab) and AI/ML group of ATS
    - The role of SFT is to provide support to the experiments on common software issues
    - Avoid duplicating efforts and focus on supporting existing activities
- Possibility to host some common ML activities shared between experiments
    - NextGen trigger activities in common ML software and algorithmic developments (not related to a specific experiment)
- Collaboration also with KT on disseminating CERN AI/ML technology and tools
- Build and maintain benchmarks and challenges (e.g. CaloChallenge for fast simulation) for testing performance of new algorithms
- Share and disseminate AI/ML knowledge: organise training in ML tools for HEP students

Interaction with existing ML activities

- We will be participating in some existing AI/ML activities in the HEP community
    - Inter-experimental ML Forum
    - Fast simulation CaloChallenge and Open Data detector
    - NextGen trigger project
- Plan to be integrated into a future general AI centre of CERN together with groups from experiments, IT, ATS and KT
    - see the discussion that happened in the Applied AI@CERN workshop

Initial Actions

- Organise ML inside SFT: bi-weekly meeting open to all SFT with reports on existing activities and possible presentation of new ideas
  - current activities in FastSim and ROOT will follow their presented plans
- Participate in initial NextGen activities on common ML topics (e.g. heterogeneous inference interface)
- Collaborate with IT (new IT ML infrastructure initiative) to implement full ML workflow for training + optimization (integrated in SWAN)
- Categorize and compare existing inference solutions for ML models
  - provide benchmarks (CPU and GPU timings vs memory usage)
- Collaborate on providing LLM for ROOT user support and code development assistant (e.g. automatic forum answers).
  - see A2rchi (MIT) and AccGPT (from CERN ATS)
- Aim to attract students in this new activity:
  - propose projects (CERN summer students) + GSOC

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