ML4EP Meeting

- Working meeting to monitor progress in current activities
  - Fast calorimeter shower simulation using ML
  - ROOT ML activities: RBatchGenerator and ML Inference (SOFIE)

- Today:
  - Introduction of GSOC students
  - New summer student
  - Status of activities and of the plan of work
    - what has been done so far

Plan of Work Presented at SFT Meeting in January

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

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.

Common SW Presentation at LHCC last week

from P. Buncic presentation at LHCC

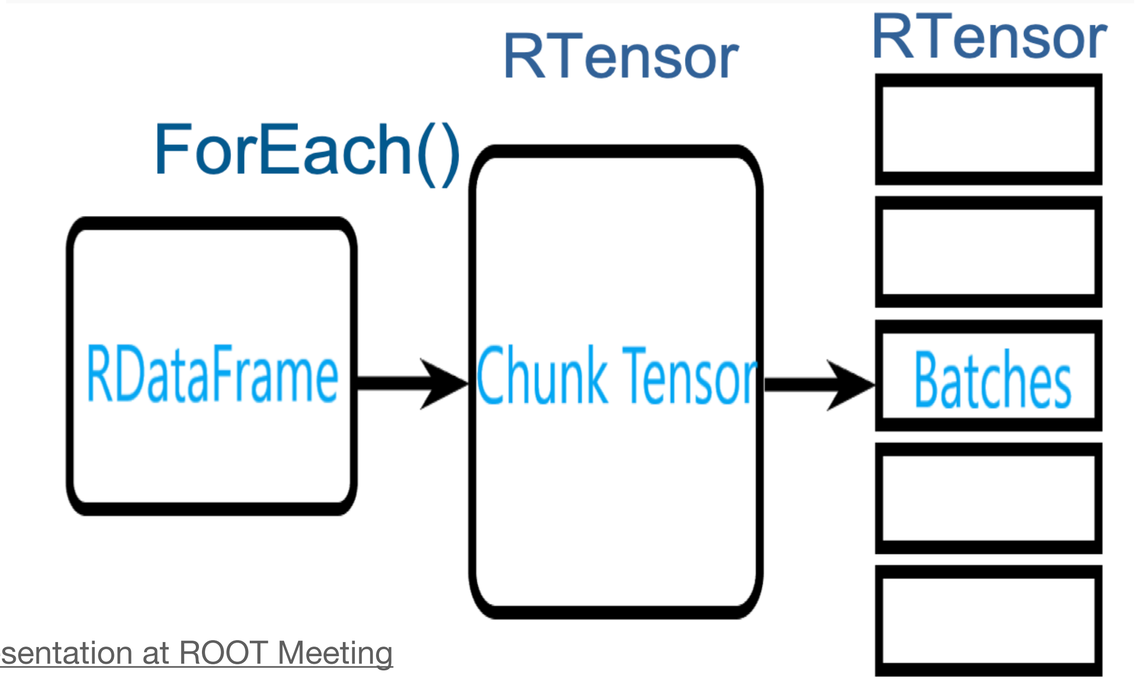
ML4EP: Plans

- **Current activities and plans for near future**
  - Validation of diffusion model (based on transformer) for ATLAS and LHCb shower simulations.
  - Work on inference optimization of diffusion model
  - Extending inference support in ROOT SOFIE for complex ML models (GNN, transformers)
  - Benchmark inference in terms of CPU time and memory consumption of common ML models used by experiments (VAE, GNN, diffusion, and transformer models)
    - using different implementations: SOFIE, Tensorflow XLA, ONNXRuntime and PyTorch
    - abstract submitted to CHEP2024
- **Longer term plans**
  - Will include tasks from NGT using their new resources
  - Develop interfaces to ML inference for integration in reconstruction and high level trigger
  - Develop common software framework for training and hyper-parameter optimisation of ML models
    - including hardware-aware NN training
  - Work on fast inference on FPGA and GPU for complex ML models
  - Contribute to community efforts in fast simulation
    - organisation of second CaloChallenge for algorithm benchmarks
    - integration of ML shower simulation models in FCCee detector simulation

ML in ROOT: Status of Activities

RBatchGenerator

- Aim to convert directly ROOT TTree data to formats used by ML frameworks for training:
  - Numpy array, PyTorch and Tensorflow Tensor
- Since last year (ROOT 630) a first version is available for generating batches directly from Tree
  - A student (Kristupas) worked on developing a direct interface from RDataFrame
  - Possible to filter data directly and generate chunks of desired size with the filter data



see Kristupas presentation at ROOT Meeting

Status of SOFIE

- Implementing (GSOC student Vedant) missing operator for parsing:
  - ParticleNet model from CMS
  - Diffusion step used in fast simulation
- Missing some operators:
  - TopK, Tile, ReduceSum and some trivial ones (ConstantOfShape, Equal)
- Vedant started implementing those missing ones
- Need to add full support for parametrised tensor shapes in all operators
  - The shape of the tensor is not fixed when generating the model, but is a parameter which can be different for every inference call (example/event)

RBDT

- RBDT class in TMVA for inference of BDT trees trained with xgboost
- Re-implemented using FastForest library from J. Rembser
  - available in ROOT 6.32
    - see https://github.com/root-project/root/pull/15173

Summary of Plan of Work

- Put RBatchGenerator in production (in progress: ~50% done)
- Consolidate RBDT (completed)
- Continue to add support for the ONNX operators (in progress)
- Benchmark inference in terms of CPU time and memory consumption of common ML models used by experiments (started now)