Why inference as a service in SML, CCE Phase 2?

Scaling ML entails scaling the training and inference ML models. Inference as a Service (IaaS) is a promising approach for scaling ML inference thanks to the following features:

1. Portable solution to supporting different coprocessors
2. Natively support event-level batching
3. Allow access to remote AI accelerators, like GPUs
4. Factorize out ML framework
5. Factorize out algorithm scheduling
6. Event batching

CCE SML should start to exam this approach and share findings with experiments. There are many ML-based algorithms employed in high energy physics. Which one CCE should prioritize?

Tracking as a service (TaaS)

Tracking is computationally expensive. However, conventional tracking algorithms are known to scale worse than linearly. Graph Neural Network (GNN)-based tracking algorithm deems a promising candidate for running tracking in GPUs. Studies about tracking will make a significant impact.

While examining GNN-based tracking, LHCb is already using GPUs for their online tracking with conventional tracking algorithms. ATLAS is exploring different approaches for event-level online tracking and wants to make a decision at the end of 2024. Now is a good time for CCE to start looking into TaaS.

CCE’s study on how the TaaS scales will be invaluable inputs to the next generation online tracking / data taking for the LHC experiments.