Fast inference in CMSSW with NVIDIA TensorRT

A. Di Pilato, A. Di Florio

Final scrum

Final scrum

Tasks:

- Integrate fast inference in CMSSW on GPU with **NVIDIA TensorRT**
- Test a simple TRT code for ParticleID and Energy Regression in HGCAL within TICL
- Modify the **`PatternRecognitionByCA`** inside TICL Framework to run with TensorRT instead of TensorFlow

Final scrum

Main issue: deprecated documentation & libraries (TRT samples didn’t help as well)

What we did:

- Managed to compile a simple TensorRT code inside CMSSW
- Adapted **`PatternRecognitionByCA`** to work with TensorRT
- Run a simple code to perform inference with the PID&ER Model outside CMSSW → however, it seems like the copy back to host is not working properly

Final scrum

What’s next:

- Debug the copy back to host
- Check the inference results
- Compare performance with the TensorFlow currently running in CMSSW

Day 3

Day 3 scrum

- Installed the latest release of NVIDIA TensorRT (TRT 6.0.1.5) within CMSSW environment as external: **compiles!**
- **`PatternRecognitionByCA`** inside TICL framework adapted to work with TensorRT instead of TensorFlow
- **Issues:** deprecated documentation & libraries

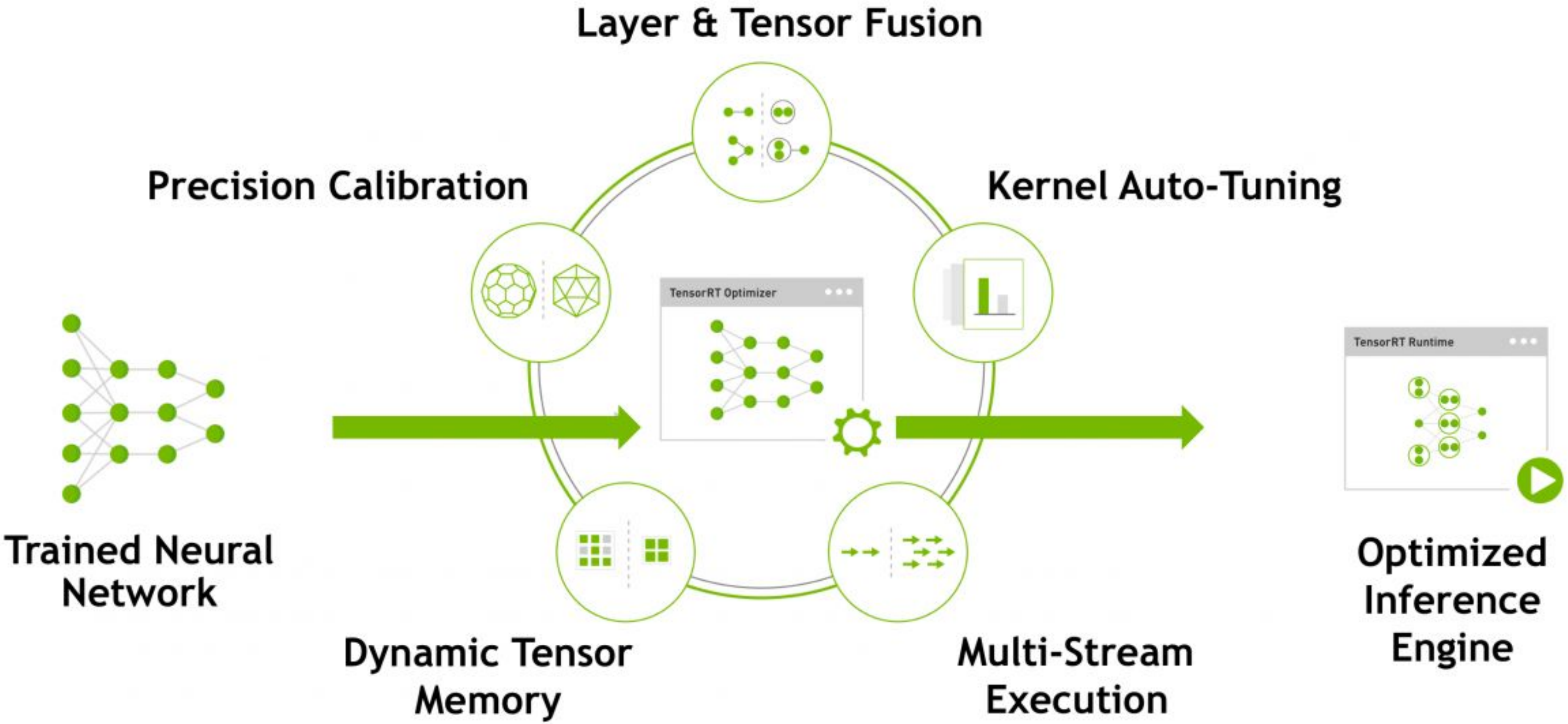
Plans for *tomorrow*: have it working in CMSSW (need to convert Tracksters to data structures)

Day 2

Day 1

Goal

- Integrate fast inference in CMSSW on GPU with **NVIDIA TensorRT**: https://developer.nvidia.com/tensorrt
- NVIDIA TensorRT is a platform for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime that delivers low latency and high-throughput for deep learning inference applications.
- Test the performance on the model developed for Particle ID and Energy Regression in HGCAL within TICL framework.



Trained Neural Network

Layer & Tensor Fusion

Precision Calibration

Kernel Auto-Tuning

Dynamic Tensor Memory

Multi-Stream Execution

Optimized Inference Engine

Day 1 scrum

- Installed the latest release of NVIDIA TensorRT (TRT 6.0.1.5) within CMSSW environment
- Verification still ongoing
  - Apparently the tool for the conversion of .pb model into .uff model is not working
  - Testing is being made on a code that worked with TRT 4.0 outside CMSSW for doublets classification in the Tracker
- A simple code will be written to work with the new problem (ParticleID and EnergyRegression in HGCAL)
- Need to adapt **`PatternRecognitionByCA`** inside TICL framework to work with TensorRT instead of TensorFlow

Day 2 scrum

- Installed the latest release of NVIDIA TensorRT (TRT 6.0.1.5) within CMSSW environment as external: **compiles!**
- Adapting **`PatternRecognitionByCA`** inside TICL framework to work with TensorRT instead of TensorFlow (need to cope with multiple outputs)
- **Issues:** deprecated documentation & libraries
    - If not mantained <u>not worth</u> to put TensorRT in production
- Plans for *tomorrow*: have a **simplified working** TensorRT example in CMSSW for PiD & energy regression