Meeting: mpp group, CERN

Syed Anwar Ul Hasan

(Postdoc fellow: Scuola Normale Superiore di Pisa)

16th September, 2021

Motivation

- Wide application of VAE, CNN VAE, Graph-based VAE models for applications
  - BSM events anomalies detection (Pratik)
  - New Physics searches for the LHC at L1 (Katya and Ema)
  - Jets-based AE for anomaly detection (Kinga)
  - New Autoencoders for L1 and HLT

Fast Inference of VAE models for HEP applications

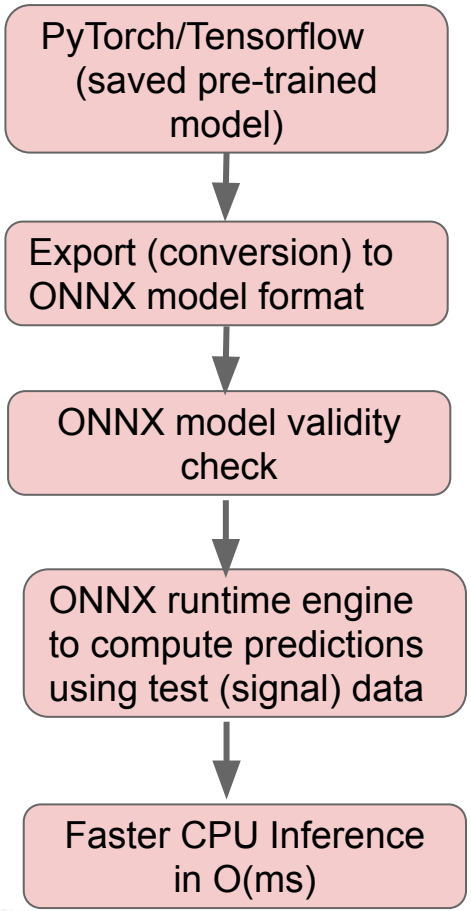
- Objective
  - Measure inference time and memory consumption on CPU, GPU, and GPU with TensorRT

- Tools:
  - For CPU Inference: Microsoft ONNX runtime engine
  - For GPU Inference: NVIDIA TensorRT with ONNX runtime

Models to consider for faster Inference

- Convolutional variational autoencoder with Jet-level representation for Anomaly detection (Kinga et. al, IML paper)
- Convolutional variational autoencoder without flows with Event-level representations for Anomaly detection (Ref: Pratik's contribution in DarkMachines Anomaly detection challenge paper)
- Other potential models: Graph-based versions of the above models, New AEs for HLT

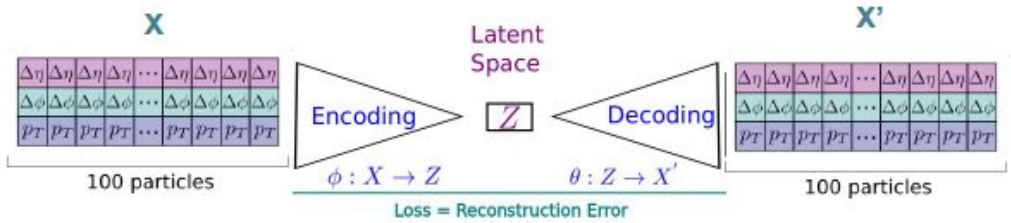
Model Conversion to ONNX format and ONNX runtime Inference



ONNX inference on CPU with Jet-level VAE - I

- Jet-level VAE model (Tensorflow)

  Input: Particle list (η, φ, pt), Jet1 & Jet2



- Model conversion to onnx model *(tf2onnx)*
- Dijet test data:
  - test_sample = [‘GtoWW15br’, ‘GtoWW15na’]

ONNX Inference on CPU with Jet-level VAE - II

- Inference results for GtoWW15br
  - Ran the onnx model with onnxruntime on test data
  - Obtain 0.52 ms inference latency with ONNX for event batch_size=1
  - CPU inference with Tensorflow around 2.28 ms for batch_size=1
  - Savings: Around 4.5 times with ONNX runtime

ONNX Inference on CPU with Event-level VAE - I

- Event-level VAE model (PyTorch)
  - Event input shape = [39,4], MET
    - 15 jets + 4 each of {bjet, m+, m-, e+, e-, gamma}
  - Number of events: 48.6K for test data (channel1)
  - Model conversion to onnx model *(torch.onnx.export)*
  - Ran the ONNX model with onnxruntime on test data samples

ONNX Inference on CPU with Event-level VAE - II

- Inference time results:
  - ONNX runtime inference time = 0.14 ms
  - PyTorch inference time = 4.11 ms
  - Savings: Around 29 times with ONNX runtime
  - Exploring further to fix this numerical inaccuracy
    - PyTorch and ONNX runtime (ORT) matching numbers differ a bit in outputs: decoded input, latent space (z)

Next Steps:

- Inference on GPU with and without TensorRT with both Jet and Event based VE models, also consider graph variants for CPU and GPU inference

- Continue with other ongoing works: Event anomaly detection with SUEPs, Image-based anomaly detection (Computer vision models)

Many thanks to **Pratik Jawahar, Kinga Anna Wozniak, and Marcel Rod** for their code and data sharing, support and research collaboration

Questions?

Number of parameters for the models

- Jet-based VAE model (vae_model.onnx)

Number of parameters: 417755 (roughly 418K)

- Event-based VAE model (test9chan1.onnx)

Number of parameters: 16356 (roughly 16K parameters)

- ParticleNet graph neural network model (particlenet_model.onnx)

Number of parameters: 587065 (roughly 587K)