Final report: Accelerating Machine Learning Reconstruction in CMS

Dr. Ruchi Chudasama, Prof. Sergei Gleyzer, University of Alabama

1 Brief Overview

The main aim of the project was the machine learning applications for physics analysis, detector reconstruction, and Phase-II upgrades, with a strong emphasis on the computational and innovative machine-learning aspects of this research. The project focused on the integration of machine learning and hardware acceleration into CMS reconstruction. It was closely tied to US-CMS Software and Computing Operations and R&D efforts in the area of heterogeneous computing resources, detector reconstruction, and analysis facilities and tools. The project aimed at developing end-to-end machine learning reconstruction and classification algorithms for the CMS data. The approach focuses on low-level detector information. The graphs and images can be formed using low-level detector information.

The specific aims of the projects were

- Extend the end-to-end deep(E2E) learning reconstruction to electron and tau identification. Integrate the electron and tau benchmarks into CMSSW E2E prototype and perform scaling studies for training and inference on heterogeneous hardware platforms relying on containerization.
- Further extend the end-to-end deep learning tau reconstruction with graph neural networks and vision and graph transformers for a combination of low-level tracker and calorimeter inputs and integrate them into CMSSW. Additionally, extend to E2E regression tasks. Perform inference and training scaling tests with heterogeneous hardware relying on containerization.

2 Techniques and outcome

The particle reconstruction algorithm at CMS experiment called Particle Flow converts detector-level information to physics objects, such as electrons, photons, muons, and hadrons, and further uses this information to build jets. Due to its capability to significantly reduce the size and complexity of particle physics data while offering a physically intuitive and simple-to-use representation for physics analyses, the Particle Flow algorithm is widely used by the ATLAS and CMS experiments. Despite the very high reconstruction efficiency of PF algorithms, some physics objects may fail to be reconstructed, are reconstructed imperfectly, or exist as fakes and limit the search beyond the standard model (BSM). Therefore it is advantageous to consider reconstruction that allows a direct application of machine learning algorithms to low-level data in the detector.

The end-to-end deep learning technique is based on high-fidelity Monte Carlo simulated event samples. The samples are produced privately for the 2018 proton-proton collisions data-taking periods at the center of mass energy 13 TeV.

Events from the production of Z/γ^* also called as drell-yan process are generated with POWHEG v2.0 at NLO in perturbative QCD and are required to decay to hadronic tau. The background processes such as W bosons in association with jets (W+jets) are generated with Madgraph@NLO. Events from top quark-antiquark pair ($t\bar{t}$) and single top quark production are generated with POWHEG v2.0 at NLO in perturbative QCD. Events from multijet production via the strong interaction, referred to as QCD multijet production, are generated at LO using PYTHIA 8.223.

The Drell-Yan process with decay to electron-positron pair is considered a signal for electron identification, is produced using POWHEG, and hadronized using Pythia8. We consider a simulated background sample of QCD dijet production produced using Pythia 8, filters are applied to select events containing electromagnetically enriched jets.

We additionally store the low-level tracker detector information, specifically, the reconstructed hits from the pixel, silicon strip detectors, and hadronic calorimeter. The most challenging part of the MC production was to obtain the correct HCAL and low-level tracker reconstructed hits information. Storing this collection at AOD step required starting the MC production from scratch and also occupied several terabytes of space.

The end-to-end deep learning approach combines a low-level detector representation and deep learning algorithms. This approach has achieved current state-of-the-art performance in identifying electrons, photons, jets, and boosted objects [1–4] using various deep-learning architecture such as convolutional neural network and graph neural network. According to the deep learning architecture to be utilized for training E2E algorithm, either **images or graphs are constructed utilizing low-level detector** information from each subdetectors of the CMS experiment. The E2E approach has shown promising results with ECAL, HCAL, and tracking hits for improved jet and boosted jet identification as well as additional information from missing energy and event vertex used for tau identification, and mass regression. Figure1 shows an end-to-end image featuring all the image layers considered in this work for a single-jet for DY→ $\tau^+\tau^-$ and top quark-antiquark pair ($t\bar{t}$) jet. Figure2 shows an end-to-end image featuring all the image layers considered in this work for DY→ e^+e^- and QCD EM-enriched jets.

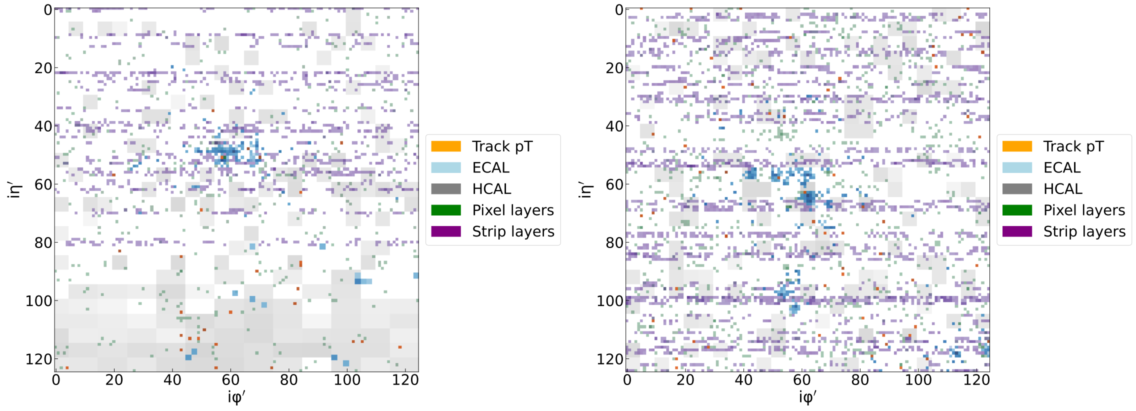

Figure 1: Composited end-to-end image of a DY→ $\tau^+\tau^-$ and top quark-antiquark pair ($t\bar{t}$) jet.

Table 1: E2E Tau Classification with EfficientNet-B0 Convolutional Neural Networks

E2E Layer Combination	AUC(CNN)
Track Pt+d0+dz+ECAL+HCAL	0.9757
Track Pt+d0+dz+ECAL+HCAL+BPIX(1-3)	0.9792
Track Pt+d0+dz+ECAL+HCAL+BPIX(1-4)+TIB/TOB1-2	0.9805

During the first phase of US-CMS SC *R&D* project, four distinct E2E benchmarks were developed: e/*gamma* classifier using information from the ECAL subdetector, quark-gluon classifier that relies on ECAL, HCAL and track p_T information, and a top jet classifier that uses hits from barrel pixel layers, track p_T, ECAL, and HCAL information. All three end-to-end classifiers use ResNet

CNNs. In another phase of the project, we developed the first **end-to-end tau identification** and **end-to-end electron vs jet identification** benchmark using end-to-end efficient Net CNNs. On top of that, we also added information from the silicon strip layers RecHits for the first time.

As can be seen in Table 1, newly developed E2E CNN EfficientNet-B0 architectures perform very well on the tau classification task for all layer combinations, and show improvements with the inclusion of additional layer information. We plan to add more input information such as secondary vertex, and missing transverse energy, and enhance and further customize the models.

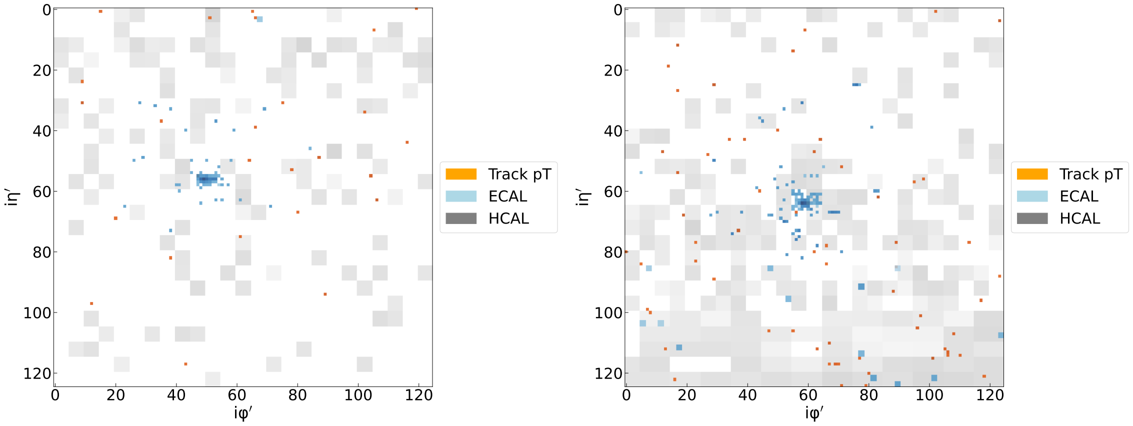

Figure 2: Composited end-to-end image of a DY→ e^+e^- and QCD EM-enriched jet.

Table 2: E2E electron Classification with EfficientNet Convolutional Neural Networks

E2E Layer Combination	Efficient Net B0	Efficient Net B3	Efficient Net B5
Track Pt+d0+dz+ECAL+HCAL	0.9980	0.9982	0.9984

As can be seen in Table 2, newly developed E2E CNN efficient Net architectures perform very well on the electron classification task even without adding the information from tracker layer RecHits.

In another part of the project, we integrated the E2E framework on top of CMSSW base classes, with a highly modular design to support customizable E2E workflows to support ML training and inference [5]. The current version can handle *fully connected*, *convolutional*, *graph neural network*, and *vision transfomer* architectures, while future versions will use other models as well. The framework consists of three main package categories: DataFormats, FrameProducers, and Taggers, as shown in Figure 3b. The DataFormats package consists of all the objects and classes needed for running the E2E modules and for storing any output back into EDM-format files. We integrated all four benchmarks in the CMSSW. Finally, with the CMSSW ONNX C++ API, we performed inference on CPU and GPU architectures within CMSSW. We considered simpleNet CNN and vision transformer architecture to benchmark the inference for all four taggers. The

Longer-term vision includes integrating end-to-end graph neural networks and transformer architectures for HGCAL reconstruction, due to their natural ability to incorporate irregular HGCAL geometry and timing inputs for a deep spatio-temporal combination with the eye towards addressing Phase-2 HL-LHC reconstruction challenges.

A part of this work on end-to-end inference within the CMS software framework was presented at the 26th International Conference on Computing in High Energy and Nuclear Physics [6].

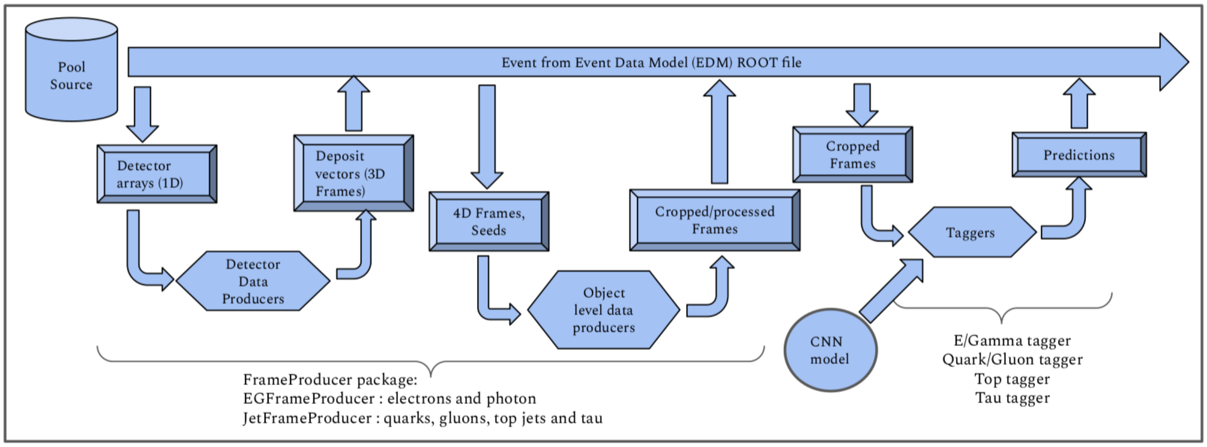

Figure 3: The end-to-end framework (`E2EFW`) pipeline used for `E/Gamma,Quark/Gluon,Top,Tau` taggers

.

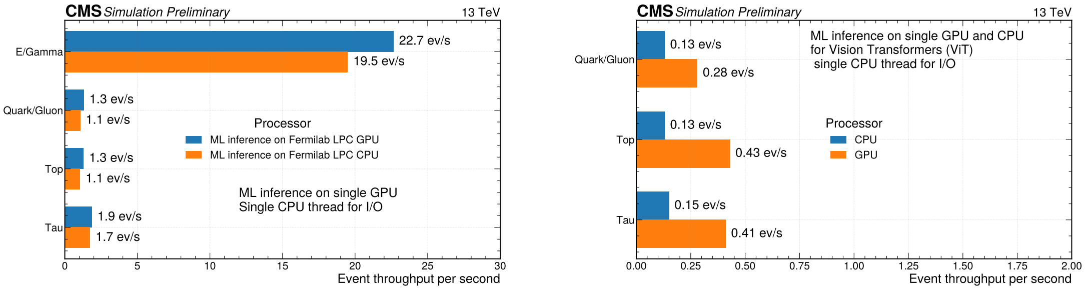

Figure 4: End-to-end inference framework event throughput per second for `E/Gamma, Quark/Gluon, Top, and Tau taggers` compared for Fermilab LPC GPU (blue) and CPU (orange) for simpleNet architecture on the left figure and vision transformer on the right figure.

References

[1] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos, *End-to-End Physics Event Classification with CMS Open Data: Applying Image-Based Deep Learning to Detector Data for the Direct Classification of Collision Events at the LHC*, *Comput. Softw. Big Sci.* **4** (2020), no. 1 6, [arXiv:1807.1191].

[2] M. Andrews, J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczos, and E. Usai, *End-to-end jet classification of quarks and gluons with the CMS Open Data*, *Nucl. Instrum. Meth. A* **977** (2020) 164304, [arXiv:1902.0827].

[3] M. Andrews *et. al.*, *End-to-end jet classification of boosted top quarks with the CMS open data*, *EPJ Web Conf.* **251** (2021) 04030, [arXiv:2104.1465].

[4] A. Hariri, D. Dyachkova, and S. Gleyzer, *Graph Generative Models for Fast Detector Simulations in High Energy Physics*, arXiv:2104.0172.

[5] **CMS** Collaboration, *End-to-end Deep Learning Inference in CMS software framework*, *CERN-CMS-DP* (2023) 036.

[6] **CMS** Collaboration, P. Chaudhari, S. Chaudhari, R. Chudasama, and S. Gleyzer, *End-to-end deep learning inference with CMSSW via ONNX using docker*, in *26th International Conference on Computing in High Energy & Nuclear Physics*, 9, 2023. arXiv:2309.1425.