Accelerating Machine Learning Reconstruction in CMS

PI: Prof. Sergei Gleyzer, Postdoc: Dr. Davide di Croce, University of Alabama

1 Introduction

The CMS experiment [1] at the Large Hadron Collider (LHC) is exploring the frontier of particle physics with highest energy proton-proton collisions ever recorded in the laboratory. The University of Alabama (UA) has been a member institution of the CMS Collaboration since 2011. The UA CMS group is led by faculty members: Sergei Gleyzer, Conor Henderson and Paolo Rumerio. Since August 2020, Professor Gleyzer supervises a postdoctoral researcher, Dr. Davide di Croce, on a US-CMS HL-LHC Software and Computing R&D project titled: *"Accelerating Deep Learning Reconstruction for CMS"*. The focus of this project is on integrating innovative machine learning algorithms into detector reconstruction, while leveraging heterogeneous computing architectures.

In this **renewal** proposal, we extend the original project to additional machine learning reconstruction targets, including applications of **graph neural networks** and **end-to-end tau reconstruction**, and further integration of machine learning and hardware acceleration into CMSSW. Detailed research plans are described in Sections §2 - §4. Sections §2.1 - §3 describe the results obtained in 2020 and the planned timeline and deliverables are discussed in Sections §4.

2 Machine Learning for New Physics at the LHC and HL-LHC

Our ability to take full advantage of machine learning during HL-LHC crucially depends on **deep integration** of information from various available sources and implementation of powerful algorithms that best exploit *highly-dimensional* and *highly-granular* data. Despite rapid recent advances, significant challenges still remain for large-scale application of **deep-learning algorithms** beyond the present state of the art. These challenges include model complexity and latency at inference time, **integration of deep learning algorithms in detector reconstruction** and simulation, the *non-euclidean nature* of some input data and *irregular geometry* posing challenges to direct application of computer vision-inspired algorithms to full detector data. These challenges and the program to address them are described in detail in the Community White Paper for Machine Learning in High-Energy Physics [2], co-edited by **PI Gleyzer**.

The LHC will enter its 'High-Luminosity' (HL-LHC) phase in 2027. The center-of-mass energy will be 14 TeV and instantaneous luminosity is projected to reach a peak of up to 7.5×10^{34} cm^{-2}s^{-1}, a factor of four increase beyond Run 2. The goal of HL-LHC is to collect an integrated luminosity of at least 3000 fb^{-1} in ten years of operations [3]. At HL-LHC, the number of pileup collisions is expected to reach 200, a significant increase above the Run 2 mean of ~ 40. To handle this challenge, and to cope with the high radiation environment, the CMS experiment is preparing major detector upgrades for HL-LHC [4]. One of the key HL-LHC challenges is how to best take advantage of these new detectors and additional advantages offered by their expanded *granularity*, *sensitivity* and *timing resolution* to enhance the physics reach and potential of the CMS experiment.

2.1 End-to-end Deep Learning for LHC and HL-LHC

For the past decade, the CMS experiment uses **particle flow** (PF) algorithms to reconstruct the LHC collision events [5]. The CMS PF algorithm attempts to *reconstruct all stable particles* in an event including muons, electrons, photons and hadrons and uses that information to build jets.

The algorithm **combines information from all CMS subdetectors** to construct PF candidates from overlapping tracks and calorimeter clusters. Currently, *particle flow* candidates are used as input for many **machine learning applications** at the LHC [6–8]. Alternatively, and crucially for future progress, *machine learning algorithms* can be **directly applied to *low-level data*** to improve the performance of particle flow algorithms, leading to a possible next-generation LHC and HL-LHC reconstruction, enhanced by machine learning techniques.

PI Gleyzer leads the development of a novel deep learning reconstruction technique suitable for LHC and HL-LHC - **end-to-end deep learning** (E2E) [9–14]. *End-to-end deep learning* leverages a combination of *deep learning* and *low-level detector representation* to efficiently identify particles and perform event-level reconstruction. This approach has achieved current *state-of-the-art performance* in identifying *electrons*, *photons*, *jets* and *boosted objects* [9–11]. End-to-end deep learning takes advantage of various deep learning architectures including *convolutional neural networks* such as ResNet [15], and *graph neural networks* that show excellent *scalability* and *generalization* to particle identification tasks [9, 16–18, 24]. Recently, PI Gleyzer has developed **graph neural network** applications for CMS Phase 2 reconstruction and *end-to-end* deep learning reconstruction and simulation [16–18]. Additional end-to-end developments include incorporation of *low-level tracking* information for improved jet and boosted jet identification, *muon* and *tau identification*, *mass regression* and *event-level anomaly detection*. Beyond standalone particle and event identification tasks, end-to-end deep learning forms the basis of an on-going CMS search for *light pseudoscalar decays* of the Higgs boson into four photons, a proof-of-concept physics application that demonstrates the ability of *innovative end-to-end deep learning* algorithms to push the experimental sensitivity in **challenging reconstruction** conditions such as fully-merged particles. Machine learning algorithms developed as part of the E2E project (such as end-to-end mass regression [19] and anomaly detection) have additional applications to novel physics searches and advanced data quality monitoring applications, that the E2E group plans to expand during 2021.

The **end-to-end** deep learning project has evolved into a major (US) CMS machine learning effort led by University of Alabama, with participating US institutions: University of Alabama, Carnegie Mellon University and Brown University, producing results both internal and external to CMS. **PI Gleyzer leads the overall E2E project**, its *algorithmic development* and *applications to reconstruction and physics analysis*, including e/γ and tau identification, anomaly detection and BSM physics searches in boosted topologies.

3 Accelerating Deep Learning Reconstruction in CMS

One of the key steps in **integrating machine learning algorithms** into CMS reconstruction is *integration of promising particle identification* algorithms such as end-to-end deep learning into the CMS Software Framework. **PI Gleyzer** and **Dr. DiCroce** have focused on integration of the end-to-end framework into CMSSW and the **inference and scaling performance** of the end-to-end algorithm described in Section §2.1. One important aspect of this work is the *understanding and optimization* of **inference time** of the E2E algorithm and its scaling with the use of heterogeneous computing and *hardware acceleration devices*, such as GPUs.

During the US-CMS SC *R&D* 2020 project, three 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 (Figure 1a). All three end-to-end classifiers use ResNet CNNs. In another phase 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

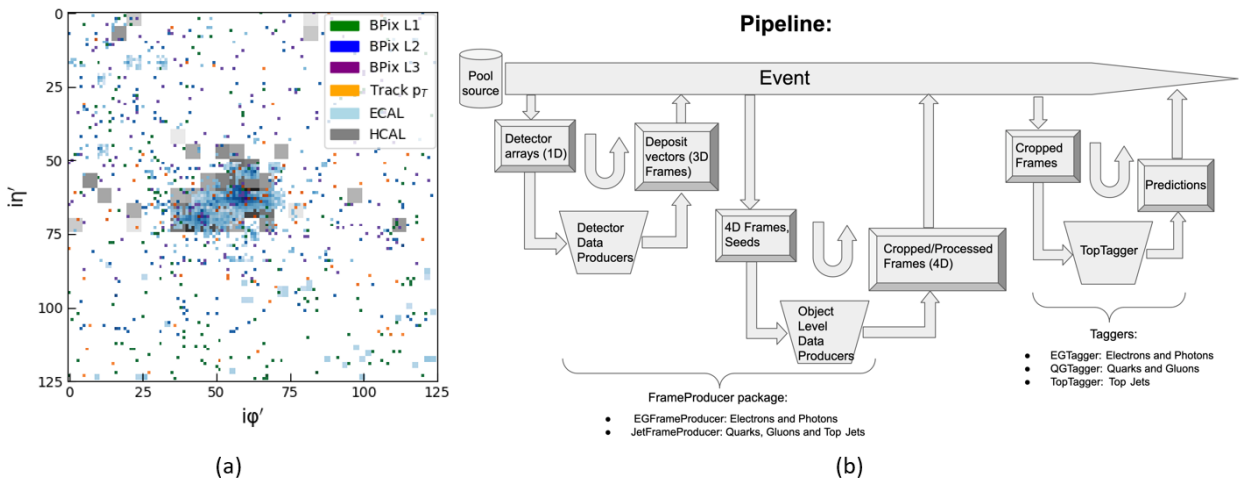

Figure 1: (a) A 6-channel composited end-to-end image of a simulated top quark jet. (b) Outline of the E2E pipeline for top classification

training and inference [12]. The current version can handle fully connected and CNN architectures, while future versions will use graph models as well. The framework consists of three main package categories: DataFormats, FrameProducers, and Taggers, as shown in Figure 1b. 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.

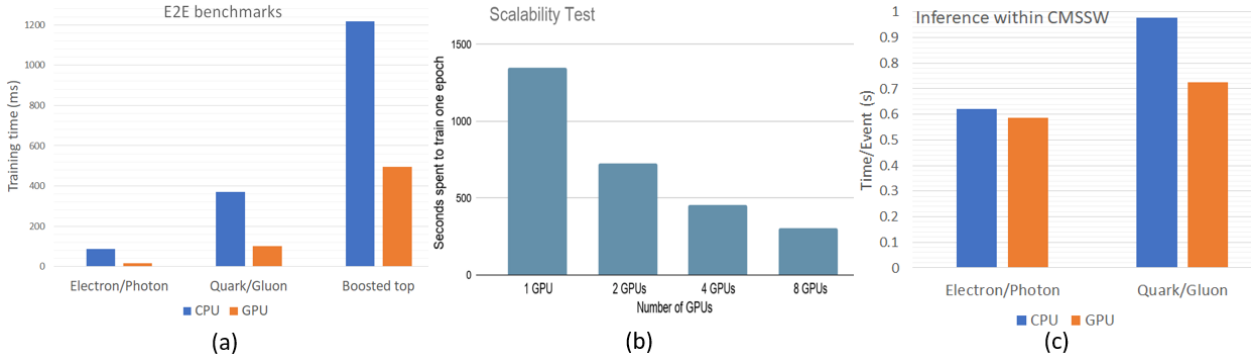

Figure 2: (a) Comparison of training time for different E2E benchmarks. (b) Scaling multi-GPU training for the standalone boosted top jet benchmark. (c) Comparison of the inference time for electron/photon and quark/gluon benchmarks using CPU and GPU architectures within CMSSW.

We also trained the E2E benchmarks on CPUs and GPUs, as shown in Figure 2a. We additionally leveraged the Nvidia Horovod framework to train E2E benchmarks on multiple GPUs and studied its scaling performance. With this procedure different layers of the deep learning model are trained on different devices, taking advantage of the inter-GPU and inter-node communication methods such as NCCL and MPI to distribute the deep learning model parameters between various workers and aggregate them. Our results show good scaling performance (Fig. 2b.). Finally, with the CMSSW TensorFlow C++ API, we performed inference on CPU and GPU architectures within CMSSW (Figure 2.c). The project has achieved all of its current milestones, and Dr. DiCroce is slightly ahead of schedule in completing the remaining milestones.

4 Research Agenda: Timeline and Milestones

Leveraging the state-of-the-art jet and boosted jet identification results achieved by the *end-to-end* deep learning algorithm that exploiting low-level tracking [10, 11], we propose to extend the E2E algorithm to **tau identification** in 2021. The goal of this work is to **improve identification of taus** across the *full momentum range*. Present deep learning techniques used for tau identification are used more traditionally with combinations of hand-selected high-level features and particle flow inputs. The *end-to-end* approach uses full detector information and deep learning algorithms applied at a much earlier reconstruction stage, and can potentially recover information losses in tau reconstruction.

- **Specific Reconstruction Aim 1**. Extend the **end-to-end deep learning** reconstruction to tau identification in 2021 with a focus on the full tau momentum range, including the low momentum regime. Integrate the tau benchmark into CMSSW E2E prototype and perform scaling studies for training and inference on heterogeneous hardware platforms.

Graph neural network models provide a **natural representation** for encoding *relational information* of physical systems. They were first proposed by Scarcelli in 2009 [20] and further developed to learn across *graph nodes* and *edges* [21]. In contrast to existing methods, *graph representation learning* can handle irregular grids with non-Euclidean geometry [22], encode physics knowledge via graph construction [23] and introduce *relational inductive bias* into data-driven learning systems [24]. One key advantage of *graph representation learning* is that the order of the objects can be permuted. **Graph representation learning** has shown early promise for LHC applications and elsewhere in particle physics [7, 16, 25–27]. Progress obtained by these approaches highlight the potential of **end-to-end deep learning algorithms with graph neural networks** for improving LHC reconstruction with upgraded CMS detectors. For example, **PI Gleyzer** has recently developed end-to-end graph generative models to successfully learn end-to-end detector representation with graph models [16–18].

- **Specific Reconstruction Aim 2**. Extend the **end-to-end deep learning** tau reconstruction to incorporate graph neural networks for combination of low-level tracker and calorimeter inputs and integrate it into CMSSW. Perform inference and training scaling tests with heterogeneous hardware.

The *end-to-end deep learning* application is an excellent test case for machine learning integration as it uses state-of-the-art machine learning algorithms and significant input space and data complexity. We will continue to build upon the initial E2E CMSSW integration in 2021 by incorporating new **reconstruction targets** and models and make it flexible enough to handle **graph inputs** at training and inference stage. Tau reconstruction offers an excellent benchmark to study both training and inference of deep learning models within CMSSW. We will continue expanding deep learning training and inference tests with E2E benchmarks on heterogeneous hardware in 2021, with a greater focus on inference and adding additional hardware architectures (FPGAs). Longer-term vision includes integrating *end-to-end graph neural networks* for HGCAL tau 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 further Phase-2 HL-LHC reconstruction challenges with end-to-end deep learning.

Period	LHC Status	Deliverables
8/21-10/21	LHC Restart	Develop standalone E2E CNN τ benchmark (BM) on single CPU and GPU Develop single-GPU and FPGA inference on E2E CNN τ BM
11/21-1/22	LHC Commissioning	Develop graph neural network (GNN) application for E2E τ BMs Perform scaling multi-GPU studies on standalone τ BMs Integrate E2E CNN τ BM into CMSSW
2/21-4/21	Run 3	Develop single-GPU inference on E2E GNN τ BM within CMSSW Integrate E2E GNN τ BM into CMSSW Benchmark E2E CNN/GNN τ CPU inference within CMSSW
5/21-8/21	Run 3	Finish E2E CMSSW integration Study E2E τ inference within CMSSW on GPU and FPGA Develop CMSSW multi-GPU inference on E2E CNN/GNN τ BMs

Milestones:

- Q3 2021. Single CPU/GPU and FPGA inference and classification performance of standalone E2E τ benchmark (BM).
- Q4 2021. GNN application for E2E on the τ BM. Multi-GPU training and inference of standalone E2E τ BM. Inference of E2E CNN τ BM within CMSSW.
- Q1 2022. Inference of E2E GNN τ BM within CMSSW. Benchmarking and performance plots of CPU inference on E2E CNN/GNN τ BM within CMSSW.
- Q2 2022. Benchmarking and performance plots of single-GPU and FPGA inference on E2E CNN/GNN τ BM within CMSSW. Multi-GPU inference on E2E CNN/GNN τ BM within CMSSW.

Project Synergy within US-CMS and CMS Machine Learning Group. The proposed research program is closely aligned with the vision and priorities of the CMS ML group. **PI Gleyzer** will foster involvement of other interested CMS groups in this key machine learning reconstruction activity, with University of Alabama providing a key push. The proposed project is also closely aligned with US-CMS Software and Computing Operations and *R&D* efforts in the area of *heterogeneous computing resources* and *detector reconstruction*. The PI will enable additional support for the proposed research program at the LPC, through collaboration with existing synergistic LPC machine learning efforts and with the broader CMS Machine Learning community.

In addition to Dr. Di Croce, UA CMS group can commit additional resources: PhD students Ana Maria Slivar and Colin Crovella to this long-term broad machine learning reconstruction integration effort. Due to the maturity of the E2E effort and its central place in CMS Machine Learning group activities, we foresee that the project will have a strong support structure to ensure long-term project success both in 2021-2022 and beyond.

5 Summary

In summary, we propose to work on integrating end-to-end deep learning algorithms into CMS reconstruction, including graph neural networks, with a focus on accelerating deep learning-enabled reconstruction algorithms for LHC Run 3 and HL-LHC, enabled by heterogeneous computing.

References

[1] CMS collaboration, *The CMS experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[2] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, *arXiv e-prints* (2018) [1807.02876].

[3] G. Apollinari, I. Béjar Alonso, O. Brüning, M. Lamont and L. Rossi, *High-Luminosity Large Hadron Collider (HL-LHC) : Preliminary Design Report*, .

[4] CMS Collaboration, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, Tech. Rep. CERN-LHCC-2015-010.

[5] CMS collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, *JINST* **12** (2017) P1003.

[6] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, *arXiv e-prints* (2017) [1702.00748].

[7] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende and K. Kavukcuoglu, *Interaction Networks for Learning about Objects, Relations and Physics*, *ArXiv e-prints* (2016) [1612.00222].

[8] M. Stoye, J. Kieseler, H. Qu, L. Gouskos and M. Verzetti, "DeepJet: Generic physics object based jet multiclass classification for LHC experiments." NIPS Workshop on Deep Learning for Physical Sciences(2017).

[9] M. Andrews, M. Paulini, S. Gleyzer and B. Poczos, *End-to-End Physics Event Classification with the CMS Open Data: Applying Image-based Deep Learning on Detector Data to Directly Classify Collision Events at the LHC*, *Computing and Software for Big Science 1* (2020) [1807.11916].

[10] M. Andrews, J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer et al., *End-to-End Jet Classification of Quarks and Gluons with the CMS Open Data*, *Nuclear Instruments and Methods A 977* (2020) [1902.08276].

[11] M. Andrews, B. Burkle, D. DiCroce, S. Gleyzer, U. Heintz, M Narain, M. Paulini, E. Usai, *End-to-End Jet Classification of Boosted Top Quarks with the CMS Open Data*, *Submitted to 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), 2021*

[12] M. Andrews, B. Burkle, S. Chaudhary, D. DiCroce, S. Gleyzer, U. Heintz, M Narain, M. Paulini, E. Usai, *Integration, Training and Inference Scaling of the End-to-End Deep Learning Framework in CMS*, *Submitted to 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), 2021*

[13] CMS Collaboration, *Exploring End-to-end Deep Learning Applications for Event Classification at CMS*, in *Proceedings, 23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018): Sofia, Bulgaria, July 9-13, 2018*, vol. 214, p. 06031, 2019, DOI.

[14] J. Alison, S. An, P. Bryant, B. Burkle, S. Gleyzer, M. Narain et al., *End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data*, in *Meeting of the Division of Particles and Fields of the American Physical Society (DPF2019) Boston, Massachusetts, July 29-August 2, 2019*, 2019, 1910.07029.

[15] K. He, X. Zhang, S. Ren and J. Sun, *Deep residual learning for image recognition*, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016, DOI.

[16] A. Hariri, S. Gleyzer,D. Dyachkova, M. Awad, D. Morozova, *Graph Generative Models for Fast Detector Simulations in Particle Physics*, in *Third Machine Learning for Physical Sciences Workshop at NeurIPS2020*, 2020, PDF.

[17] A. Hariri, D. Dyachkova, S. Gleyzer, *Graph Generative Neural Networks for Fast Detector Simulations in High-Energy Physics*, *Submitted to 25th International Conference on Computing in High Energy and Nuclear Physics (CHEP2021), 2021*

[18] A. Hariri,D. Dyachkova, S. Gleyzer, *Scaling Graph Generative Models for Fast Detector Simulations in High Energy Physics*, *to appear in Nvidia GPU Technology Conference (GTC2021), April 11-16, 2021*

[19] CMS Collaboration, *Reconstructing Unresolved Particle Decays via Deep Learning with Analytic Continuation*, AN-20-149/EGM-20-001, 2021

[20] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, *The graph neural network model*, *IEEE Transactions on Neural Networks* **20** (2009) 61.

[21] Y. Li, D. Tarlow, M. Brockschmidt and R. Zemel, *Gated Graph Sequence Neural Networks*, *arXiv e-prints* (2015) [1511.05493].

[22] J. Bruna, W. Zaremba, A. Szlam and Y. LeCun, *Spectral networks and locally connected networks on graphs*, 1312.6203.

[23] D. Zheng, V. Luo, J. Wu and J. B. Tenenbaum, *Unsupervised learning of latent physical properties using perception-prediction networks*, 1807.09244.

[24] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski et al., *Relational inductive biases, deep learning, and graph networks*, 1806.01261.

[25] X. Ju et al., *Graph Neural Networks for Particle Reconstruction in High Energy Physics detectors*, 2003.11603

[26] N. Choma et al, *Track Seeding and Labelling with Embedded-space Graph Neural Networks*, 2007.00149

[27] A. Heintz et al, *Accelerated Charged Particle Tracking with Graph Neural Networks on FPGAs* 2012.01563