IRIS-HEP Fellowship Proposal: GPU-accelerated Machine-learned Particle-flow Reconstruction

Farouk Mokhtar, UC San Diego
Duration: March 29, 2021 – September 10, 2021 (50% FTE)
Area: Innovative Algorithms

1 Proposal

Machine learning (ML) has been used in many facets of LHC computing and physics analyses and provides an alternative paradigm for designing reconstruction algorithms. However, the main uses of these algorithms have been at a high-level in the analysis after reconstruction has been performed with traditional algorithms. A greater potential gain may be found in developing ML algorithms that perform event reconstruction based on rich, lower-level detector information. An important core software algorithm is the particle-flow (PF) reconstruction algorithm [1–3], which takes disparate types of tracks and clusters reconstructed from different subdetectors and returns a list of final-state PF candidates.

I propose to apply state-of-the-art ML techniques, mainly graph neural networks (GNNs), to learn from non-Euclidean structured data, to the task of PF reconstruction in CMS and for LHC detectors more generally. GNN-based architectures have been shown to be performant for tasks ranging from charged particle tracking [4, 5], to jet classification [6–8], and clustering [5]. Thus, it is interesting to explore the possibility of improving core reconstruction algorithms with highly parallelizable models, such as GNNs, given the expected increased need for computing resources in the HL-LHC era. On the computing side, these ML models are easier to accelerate with heterogeneous computing resources, such as GPUs and FPGAs, which gives them an advantage over traditional PF algorithms. Previous work has been conducted within CMS in a collaboration between UC San Diego, the National Institute of Chemical Physics and Biophysics (NICPB) in Estonia, Caltech, and CERN [9] using a GNN to reproduce the existing PF algorithm. Two of the goals of my project would be (1) to attempt to improve upon the existing PF algorithm by training directly on generator-level "truth" particle information and (2) to develop and train on generic datasets, potentially allowing for the transferability of this work to other LHC experiments like ATLAS as well as future colliders like FCC.

This project will be conducted under the local supervision and guidance of Professor Javier Duarte (UC San Diego) with additional support from Professor Mark Neubauer (UIUC) and other IRIS-HEP members working on GNNs for event reconstruction. The task of the GNN will be to perform regression by predicting a vector of PF candidate features, including energy, and momentum, based on generator-level information. We will train on generic detector simulated data created using Delphes [10], to be made public as a benchmark dataset, as well as official CMS simulation. Different GNN models will be explored to find the ones best suited to this task. A comparison between traditional PF and ML-PF based on generator-level truth information will be made. Finally, some work will be dedicated to accelerating the ML-PF inference with GPUs using Nvidia Triton inference server, a marked advantage over the traditional PF algorithm. This work will be done in collaboration with the Fast Machine Learning Lab, which has already developed tools for this task [11–13].

2 Timeline

The timeline I propose is a total of six months (spring plus summer), during which I will work 50% of the time on this research. I plan to also take two academic graduate courses in the spring, but I will have no teaching or other research obligations during this time.

- **Month 1**: Generate Delphes and CMS datasets with appropriate particle truth definition. Make Delphes dataset public on Zenodo.
- **Month 2–3**: Explore different GNN models and evaluate their performance with respect to the several metrics, including classification accuracy and energy resolution.
- **Month 4–5**: Compare traditional PF and ML-PF based on generator-level truth information. Present work internally in CMS meetings and IRIS-HEP meetings.
- **Month 6**: Accelerate the ML-PF inference with GPUs using Nvidia Triton inference server. Ensure all code is available publicly at Ref. [9].

References

[1] ALEPH Collaboration, "Performance of the ALEPH detector at LEP", *Nucl. Instrum. Methods Phys. Res. A* **360** (1995) 481, `doi:10.1016/0168-9002(95)00138-7`.

[2] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *J. Instrum.* **12** (2017) P10003, `doi:10.1088/1748-0221/12/10/P10003`, `arXiv:1706.04965`.

[3] ATLAS Collaboration, "Jet reconstruction and performance using particle flow with the ATLAS Detector", *Eur. Phys. J. C* **77** (2017) 466, `doi:10.1140/epjc/s10052-017-5031-2`, `arXiv:1703.10485`.

[4] S. Farrell et al., "Novel deep learning methods for track reconstruction", in *4th International Workshop Connecting The Dots (CTD)*. 2018. `arXiv:1810.06111`.

[5] X. Ju et al., "Graph neural networks for particle reconstruction in high energy physics detectors", in *Machine Learning and the Physical Sciences Workshop at the 33rd Annual Conference on Neural Information Processing Systems*. 2019. `arXiv:2003.11603`.

[6] E. A. Moreno et al., "Interaction networks for the identification of boosted $H \to b\bar{b}$ decays", *Phys. Rev. D* **102** (2020) 012010, `doi:10.1103/PhysRevD.102.012010`, `arXiv:1909.12285`.

[7] E. A. Moreno et al., "JEDI-net: a jet identification algorithm based on interaction networks", *Eur. Phys. J. C* **80** (2020) 58, `doi:10.1140/epjc/s10052-020-7608-4`, `arXiv:1908.05318`.

[8] H. Qu and L. Gouskos, "ParticleNet: Jet tagging via particle clouds", *Phys. Rev. D* **101** (2020) 056019, `doi:10.1103/PhysRevD.101.056019`, `arXiv:1902.08570`.

[9] J. Pata and J. Duarte, "particleflow". [software], 2020. `https://github.com/jpata/particleflow`.

[10] DELPHES 3 Collaboration, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment", *JHEP* **02** (2014) 057, `doi:10.1007/JHEP02(2014)057`, `arXiv:1307.6346`.

[11] J. Duarte et al., "FPGA-accelerated machine learning inference as a service for particle physics computing", *Comput. Softw. Big Sci.* **3** (2019) 13, `doi:10.1007/s41781-019-0027-2`, `arXiv:1904.08986`.

[12] J. Krupa et al., "GPU coprocessors as a service for deep learning inference in high energy physics", (2020). `arXiv:2007.10359`. Submitted to *Mach. Learn.: Sci. Technol.*

[13] D. S. Rankin et al., "FPGAs-as-a-service toolkit (FaaST)", in *2020 IEEE/ACM International Workshop on Heterogeneous High-performance Reconfigurable Computing (H2RC)*. 10, 2020. `arXiv:2010.08556`.