Fast neural networks for the ATLAS L0 muon trigger

M. Errico, on behalf of the ATLAS collaboration
Università di Roma La Sapienza - Rome, Italy
INFN, Sezione di Roma - Rome, Italy

Summary. — The identification and reconstruction of muon tracks are essential for high energy physics experiments and must be performed in real time for trigger algorithms. In LHC experiments like ATLAS, triggering will become more complex with the HL-LHC upgrade, which will significantly increase luminosity. The ATLAS Muon Spectrometer (MS) uses Resistive Plate Chambers (RPCs) for real-time transverse momentum reconstruction in triggering. As part of the Phase-II upgrade, an additional RPC station will be installed to mitigate efficiency losses from gas mixture changes and enable higher acceptance coincidence schemes. Phase-II will also introduce a new trigger architecture, processing RPC data off-detector using Field-Programmable Gate Arrays (FPGAs). This study explores lightweight machine learning algorithms for the first-level (L0) barrel muon trigger, with latency on the order of O(100 ns), made possible by FPGA parallelisation. Two development tracks are presented: a Convolutional Neural Network (CNN) prototype and an ongoing Graph Neural Network (GNN) model. These approaches aim to improve the selection efficiency of muons in the MS, especially for rarer cases like non-prompt muons, by relaxing the geometrical constraints of traditional trigger algorithms.

1. – The ATLAS Muon Spectrometer

The Muon Spectrometer (MS) of the ATLAS experiment [1] at LHC is a large, instrumented volume of air with roughly cylindrical geometry around the z axis (beam axis); its barrel is divided in two sides over z and sixteen sectors over ϕ. Each sector contains three doublets of Resistive Plate Chambers (RPCs) distributed along the radial direction, covering the pseudorapidity range $|\eta| < 1.05$. RPCs are fast gaseous detectors with a nominal space-time resolution of $1\,\text{cm} \times 1\,\text{ns}$; they are independently segmented in η and ϕ and are used for triggering and ϕ coordinate measurements. The MS operates within a 0.5 T toroidal magnetic field, which bends charged particles in the (η, r) plane. This bending allows the measurement of the transverse momentum (p_T) of muons from the η positions recorded by the RPCs. These p_T measurements are used in the Level-0 (L0) barrel muon trigger to select muon candidates based on their momentum.

In preparation for the high-luminosity [2] Phase-II of the experiment, the MS barrel will undergo a significant upgrade: an additional triplet of RPCs will be installed in

the inner barrel region, increasing the total number of RPC stations per sector to four [3]. This upgrade will allow the implementation of a coincidence scheme with higher acceptance. Phase-II will also introduce a new trigger system architecture: during operation, RPC measurements will be sent off-detector to Field-Programmable Gate Arrays (FPGAs), where the first level (L0) trigger logic will be implemented [4]. Each sector's data will be handled by a Xilinx Virtex Ultrascale+ FPGA (XVU13P). The upgraded L0 sector logic will select muon candidates based on transverse momentum thresholds in the $[4, 80]$ GeV range and is designed to handle multiple candidates per half-sector per event. The required output from the sector logic will include:

- the highest p_T threshold satisfied
- the estimated p_T, η, charge (q), and ϕ of each identified muon
- the coincidence scheme satisfied
- the estimated z coordinate of the muon on each RPC station.

2. – Neural Network trigger algorithms

The Phase-II TDAQ layout is ideal for the deployment of lightweight neural networks due to the parallelisation capabilities of FPGAs. Neural network trigger algorithms based on pattern recognition offer a compelling alternative to traditional algorithms, which often require geometrical hypotheses such as the assumption that all muons come from the interaction point.

The RPC data received by the sector logic includes the η measurements for each hit in the MS which, combined with the information of the radial position of each RPC station, can be used to reconstruct each identified track. The precision of the pseudorapidity information is determined by the number of η strips in which each RPC station is segmented along z.

The dataset used for this study is comprised of events with simulated muons coming from the interaction point, uniformly distributed in $\eta \in (0, 1.05)$ and $p_T \in (3, 30)$ GeV. The geometry, magnetic field and resolution of the MS have also been simulated, although the dataset does not accurately represent the detector acceptance due to only non-empty events with hits on at least two stations having been selected. Additionally, noise events have been simulated, and the same type of noise has been added to the muon events. The noise events are based on a simplified model of the background according to which hit clusters are uniformly distributed across stations and the pseudorapidity; thus, the correlations between noise hits or between background and signal have not been simulated.

3. – Convolutional Neural Network

The first architecture evaluated is a Convolutional Neural Network (CNN), a model well-suited for image processing and pattern recognition. An initial version of this model has been developed, as described in [5].

To be processed by the CNN, RPC data must be encoded into a rectangular image grid. The pseudorapidity measurements are discretized into 384 bins (a value selected as a compromise based on the number of strips in each RPC station). Given that there are four RPC stations, the CNN input is structured as a 4×384 pixel image. Based on this input, the model predicts the transverse momentum (p_T) of the candidate muon track.

The model is compressed using Knowledge Distillation (KD), Quantization-Aware Training (QAT), and input fragmentation, resulting in approximately 1k parameters represented with 3-bit precision. Its performance, evaluated via efficiency turn-on curves around nominal p_T thresholds, is very promising. For a 10 GeV threshold, the model achieves 90% selection efficiency at 17 GeV with a background acceptance rate below 0.2%. The design is implemented using the hls4ml library [6, 7], achieving a latency of 440 ns and minimal FPGA resource usage.

The same CNN architecture has been extended to handle multi-track events and support classification outputs. Multi-track events are constructed by combining single-muon events. The dataset is balanced, with signal events equally distributed between one, two and three-track configurations. The compressed model, produced using KD and QAT, has 8.4k parameters at 6-bit precision; it predicts the number of muons and reconstructs the charge, η and p_T of up to two leading tracks. In this configuration, turn-on efficiency curves are computed separately for leading and subleading muons at the 10 GeV threshold (fig. 1). Both curves reach 90% efficiency by 14 GeV, and the total accepted background fraction remains below 0.2%. However, the efficiency for subleading muons with $p_T \in (3, 4)$ GeV is approximately 10%, indicating limited rejection of low-momentum muons in multi-track events.

4. – Graph Neural Network

The other architecture evaluated is a Graph Neural Network (GNN), using the Interaction Network model presented in [8], a simple message passing architecture operating on static graphs. Static graphs are sets of nodes and edges connecting the nodes which remain fixed throughout the model; therefore only the node and edge features are updated during evaluation, without changing the structure of the graph, resulting in a lower latency for each step of message passing. The GNN can directly process the (η, r) RPC hits as nodes of the graph. However, the graphs obtained from the RPC measurements must be processed by a graph-building algorithm which adds the connections between the nodes before being used as input for the GNN. The decision on whether or not to connect two nodes is made on a purely kinematic basis: if the (η, r) coordinated of the two nodes satisfy the relation

$$|\tanh \eta_2 - \tanh \eta_1| = \frac{|r_2 - r_1|}{R} \tag{1}$$

where the hyperparameter R is the minimum assumed curvature radius, then an edge connecting the two nodes is created, with $(|\eta_2 - \eta_1|, |r_2 - r_1|)$ as a feature vector. Edges connecting signal nodes are labelled as signal edge, while all other edges are labelled as background. Currently the model has been trained and tested only on background and single-track events.

The model can be used to classify edges or to directly predict the physical quantities of interest. In the first configuration, after three steps of message passing, each consisting of a dense block which updates the edge features, an aggregation operation on the edge features and a dense block which updates the node features, another edge update block is added with a single sigmoid output, representing the probability of each edge being a signal edge. The classification output for a single-track event is shown in fig. 2 as an example. Since the original simulated dataset does not distinguish between RPC hits left by the muon or by other related events (*e.g.* bremsstrahlung), it is possible for the

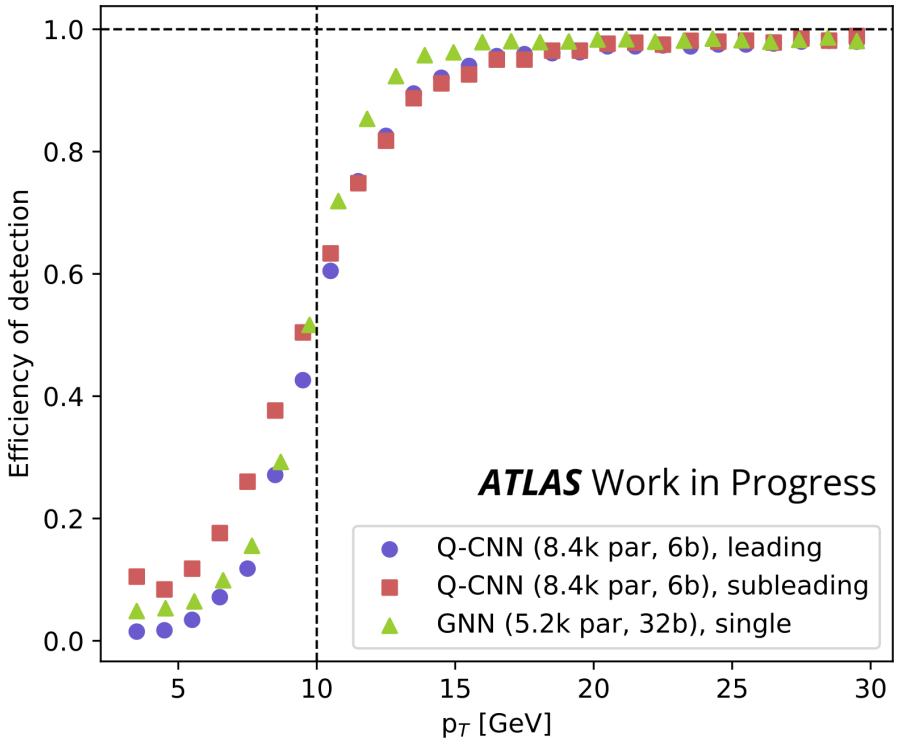

Fig. 1. – Efficiency turn-on curves for two of the models presented in this study. The violet (red) curve represents the performance of the CNN in the selection of leading (subleading) muons, while the green represents the preliminary estimate of the performance of the GNN on single-track events.

model to reject even a large fraction of signal edges in events such as the one presented, in a way that does not impede the reconstruction of the actual track.

The most direct way to estimate the ability of the model to reconstruct the event, however, is to include the reconstruction in the model itself and to train over the transverse momentum labels. This is achieved by adding an adaptive pooling operation after the three message passing blocks which extracts a single feature vector from the node and edge features of each event, and by using said vector as input to a fully connected reconstruction block. The resulting performance is once again evaluated through the efficiency turn-on curve around the 10 GeV momentum threshold. A preliminary comparison between this last architecture and the CNN one is shown in fig. 1. Although not a direct comparison, since the GNN can only handle single-track events, the performance of the GNN is promising, as it is comparable to that of the CNN on the leading muons.

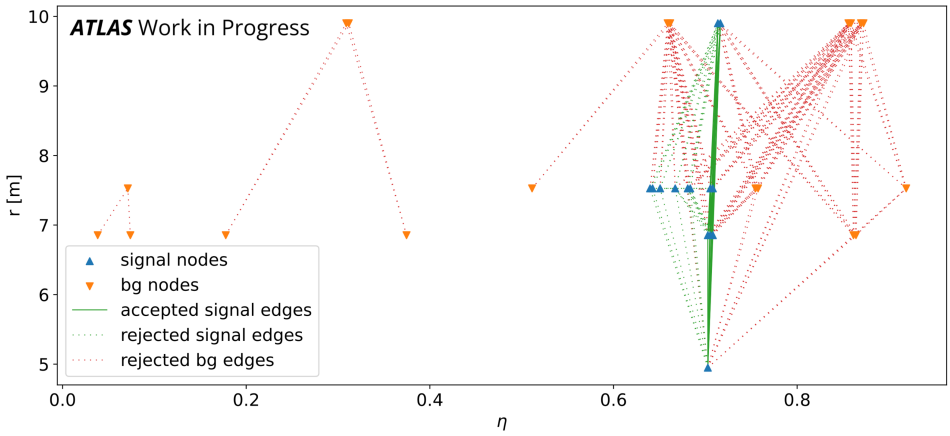

Fig. 2. – Example of the output of the GNN model when used to classify edges. The graph represents an event with $\eta = 0.706$ and $p_T = 20.8$ GeV. The signal and background edges are represented in different colours. Accepted edges are solid while rejected ones are dotted. For this event there are no accepted background edges.

5. – Conclusion

The development of fast neural networks for the ATLAS L0 muon trigger could represent a significant advancement in real-time event selection for the HL-LHC. The CNN approach demonstrates strong performance in the single-track case, achieving high efficiency and low latency within FPGA constraints, but the performance at low transverse momentum does not immediately extend to the multi-track case. The GNN architecture shows similarly promising results for single-track reconstruction, and is expected to scale well with the maximum number of muons per event due to its ability to directly examine the connectivity properties of the RPC hits.

Future work will focus on extending the GNN to handle multi-track events and classification outputs, as well as model compression (KD, QAT) and optimisation to simplify FPGA deployment. Additional improvements to the GNN architecture are possible, such as the use of a local measurement of the magnetic field as another node feature.

* * *

The author acknowledges the help of Dr. D. Fiacco, Prof. S. Giagu, Dr. G. Gustavino, Dr. V. Ippolito and Dr. G. Russo.

REFERENCES

[1] The ATLAS Collaboration, *JINST*, **3** (2008) S08003 *The ATLAS Experiment at the CERN Large Hadron Collider*
[2] Aberle O. *et al.*, CERN-2020-010 (2020) *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*
[3] The ATLAS Collaboration, CERN-LHCC-2017-017, ATLAS-TDR-026 (2017) *Technical Design Report for the Phase-II Upgrade of the ATLAS Muon Spectrometer*

[4] The ATLAS Collaboration, CERN-LHCC-2017-020, ATLAS-TDR-029 (2017) *Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System*
[5] Francescato S., Giagu S., Riti F. *et al.*, *Eur. Phys. J.*, **81** (2021) 969 *Model compression and simplification pipelines for fast deep neural network inference in FPGAs in HEP*
[6] Duarte, Javier *et al.*, *JINST*, **13** (2018) P07027 *Fast inference of deep neural networks in FPGAs for particle physics*
[7] Aarrestad, Thea *et al.*, *Mach. Learn. Sci. Tech.*, **2** (2021) 045015 *Fast convolutional neural networks on FPGAs with hls4ml*
[8] Elabd A., Razavimaleki V., Huang S. *et al.*, *Frontiers in Big Data*, **5** (2022) *Graph Neural Networks for Charged Particle Tracking on FPGAs*