PROCEEDINGS OF SCIENCE

Machine Learning for Real-Time Processing of ATLAS Liquid Argon Calorimeter Signals with FPGAs

Raphaël Bertrand[a,*] **on behalf of the ATLAS Liquid Argon Collaboration**

[a]*Aix Marseille Univ, CNRS/IN2P3, CPPM,*
Marseille, France

E-mail: bertrand@cppm.in2p3.fr

The Large Hadron Collider (LHC) will increase its instantaneous luminosity by a factor of 5-7 for the High Luminosity LHC (HL-LHC) upgrade. At HL-LHC, the number of proton–proton (p-p) collisions per bunch crossing (pileup) will rise significantly, imposing stringent requirements on detector electronics and real-time data processing. The ATLAS Liquid Argon (LAr) calorimeter, measuring the energy of electrons and photons, will be upgraded to prepare it for the high rates expected at the HL-LHC. The full LAr readout electronic chain will be replaced, namely a new off-detector board (LASP) will be installed to compute the energy deposited in the detector. The LASP board is equipped with state-of-the-art FPGAs with greater processing power and memory, enabling the deployment of more advanced algorithms to compute the energy and replace the optimal filtering (OF) algorithm. Four neural network architectures are presented, and their improvements in energy resolution compared to the legacy filter algorithms are discussed. A Bayesian optimisation of the neural networks hyperparameters is used to ensure the best performance while limiting the network size for FPGA implementation. Moreover, the Deep Evidential Regression (DER) is used to compute the uncertainty on the network prediction without significant increase in the network size.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

[*]Speaker

PoS(EPS-HEP2025)078

A. Introduction

The Large Hadron Collider (LHC) is a particle accelerator where bunches of protons (p) are accelerated to nearly the speed of light and collide at 40 MHz, which corresponds to one p–p bunch crossing (BC) every 25 ns. These collisions can reach energies of up to 14 TeV and produce highly energetic new particles at the interaction points. The ATLAS detector is located at one of these interaction points and measures the properties of the resulting particles. The Liquid Argon Calorimeter (LAr), one of ATLAS sub-detectors, measures the energy of electromagnetic particles such as electrons and photons. Each of its 182,468 cells generates an electronic signal pulse (shown in Figure 1) with an amplitude proportional to the deposited energy by the incident particles. This signal spans 625 ns (25 BC) and is subsequently shaped, sampled, and digitized at 40 MHz. Currently, samples near the peak of the pulse are used to reconstruct the deposited energy using an Optimal Filtering (OF) algorithm. By the end of 2026, the LHC will shut down for an upgrade and is scheduled to restart in 2030. The luminosity will increase by a factor of 5–7, and the maximum number of simultaneous p–p collisions will rise to 200. This high-pileup environment will affect the performance of the LAr calorimeter. To handle the increased data rate and trigger demands, the LAr readout electronics will be replaced, and a new dedicated board (LASP) has been designed to compute the energy. The LASP board will host two state-of-the-art FPGAs to implement the energy reconstruction algorithms, with each board reconstructing the energy for 384 channels. Several hundred boards will be required to cover the full calorimeter. This upgrade provides an opportunity to implement more complex algorithms, such as neural networks (NNs) to reconstruct the energy. High luminosity causes two effects that can degrade the performance of the Optimal Filtering (OF) algorithm: in-time pileup and overlapping pulses. In-time pileup occurs when energy is deposited by multiple particles during the same BC and it is not treated in this document. Overlapping pulses arise when significant energy is deposited in a preceding BC close to the current BC, causing a distortion of the pulse shape.

B. Energy reconstruction algorithms

The energy reconstruction algorithms must treat online a large flux of data : the energy must be reconstructed for each BC with a latency below 125 ns starting from the BC of the energy deposit. In addition, the algorithm must fit within the resource limits of the FPGAs. The number of Multiply-Accumulate Operations (MAC units) can be used to probe the possibility to implement algorithm on an FPGA. It has been shown that algorithms requiring fewer than 500 MAC units are suitable for implementation on the LASP FPGAs [1]. The OF algorithm is designed to compute the deposited energy using a weighted sum of five samples following the energy deposit. More complex approaches such as NNs are investigated to be put on FPGAs. The NNs are expected to exploit information from samples prior to the energy deposit to correct for pulse distortions, in addition to four samples starting from the BC of the energy deposit (shown in Figure 1).

PoS(EPS-HEP2025)078

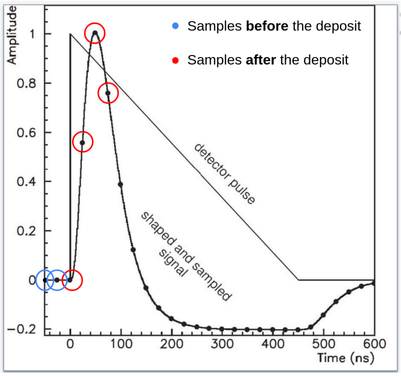

Figure 1: LAr cell electronic pulse, shaped, digitised, and sampled at 40 MHz. Samples around the pulse peak (red) are used for energy computation by the optimal filtering algorithm. The neural networks use in addition the samples taken before the energy deposit (blue) to correct for potential overlapping pulses.

Four different architectures of NNs are presented and optimised : RNN, CNN, Dense+RNN, and Dense. Preliminary studies [2] have shown that CNN and RNN are capable of correcting pulse distortions and accurately reconstruct the deposited energy, whereas the performance of OF degraded under high luminosity and overlapping pulses. The effectiveness of the correction depends on the network size, requiring optimisations to balance performance and FPGA implementation. The RNN architecture is composed of multiple cells that share the same parameter values. The input for each RNN cell is composed of the output from the preceding cell and the digitised sample at a given BC. Finally, a dense layer concatenates the output of the last cell to provide the predicted energy. The RNN architecture allows for low latency implementations since the computations can start at the arrival of the first sample. However, while large RNNs offer good performance over the full dynamic range, the number of MAC units tends to be too high for straightforward implementation on FPGAs. The CNN architecture uses kernels sliding over the full range of the input layer, producing one output neuron per slide. In addition, a unique kernel of the same size is used per filter. The CNN architecture in this study employs three convolutional layers. Two other architectures, Dense+RNN and Dense, are proposed to reduce the number of MAC units while maintaining the latency low. An optimisation of the RNN architecture (Dense+RNN) (shown in Figure 2) is proposed, limiting the RNN cells to samples after the energy deposit and using an additional dense layer to initialise the first RNN with samples prior to the energy deposit as input to the dense. This reduces the number of MAC units because dense layers scale linearly with their number of units, whereas RNN layers scale quadratically. A Dense architecture (shown in Figure 3), exclusively based on dense layers, is also proposed. A first block of dense layers is applied to the samples prior to the energy deposit to initiate the computation before the later samples arrive. Then, the output of this block is combined with the later samples to finalise the energy computation.

C. Neural networks hyperparameters tuning

In this section, hyperparameter optimisation results using Bayesian methods are shown. The number of samples prior to the energy deposit and the dimensions of the intermediate layers must be tuned. Increasing the samples after the energy deposit leads to increased latency without significant

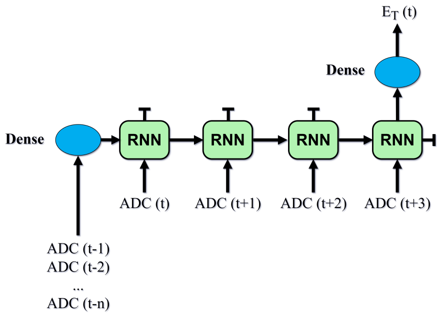

Figure 2: The Dense+RNN architecture. Information from samples prior to the energy deposit is compressed using a dense layer, which is connected to the initialization of the first RNN cell. Samples after the energy deposit are fed into four RNN cells. Finally, a dense layer concatenates the output of the last RNN cell to compute the reconstructed energy. The input samples are labeled as ADC, as they are generated by an analog-to-digital converter.

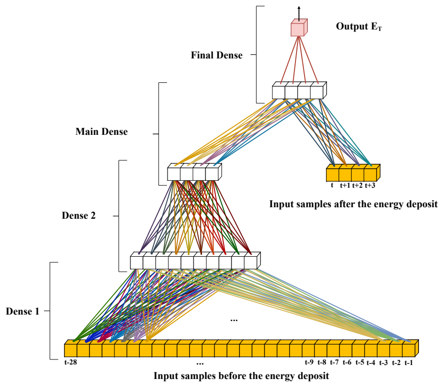

Figure 3: The Dense architecture. A first block of dense layers (Dense 1 and Dense 2) processes the samples prior to the energy deposit to initiate the computation. The later samples, combined with the output of this first block, are then fed into a main dense layer followed by the final dense, which computes the final reconstructed energy.

gain. The CNN architecture is quite different as the dimensions of the layers is determined by the kernel size, the number of filters and the dimensions of the previous convolutional layer. Thus, the kernel size and the number of filters are optimised.

Bayesian optimisation is a method for finding the best parameters to maximise a performance function while minimising the number of evaluations. This is particularly useful when function evaluation is expensive, such as in NN training. The optimisation begins by initializing several random points, in this case by training models with randomly selected hyperparameters. From this starting point, iterative refinement explores the optimal parameter space. Each iteration consists of three steps: interpolation between existing evaluations to identify correlations between parameters and performance, use of an acquisition function to select the next evaluation point, and evaluation of the performance function at the selected point.

The interpolation between points uses a Gaussian kernel to characterise a mean value and the associated uncertainty of the approximated performance function. The acquisition function balances exploration (high-uncertainty regions) and exploitation (high-mean regions). The performance function balances model performance with hardware constraints to ensure FPGA feasibility. It uses the energy resolution on a test dataset and the number of MAC units required. As described in [3], a linear hardware penalty is applied when the number of MAC units exceeds 500, and an exponential penalty beyond 850 MAC units. This constrains NN size to a practical range and balances energy resolution gains against hardware resource usage.

The RNN model shown here was taken from preliminary studies [2], as Bayesian optimisation did not converge to an improved configuration. This allows a direct comparison and highlights the gains achieved with the optimised architectures relative to these early results. Figure 4 shows the energy scale represented by the mean of the difference between the predicted transverse energy ($E_{\mathrm{T}}^{\mathrm{pred}}$) and the true transverse energy ($E_{\mathrm{T}}^{\mathrm{true}}$) as a function of the true transverse energy. $E_{\mathrm{T}}^{\mathrm{true}}$ corresponds to the sum of in-time pileup and a hard scatter event corresponding to an injecting flat energy from 0 to 130 GeV (as described in [3]). The Dense+RNN, Dense, and CNN architectures provide a better energy scale compared to RNN and OF. This behaviour is expected for the OF, which is designed to estimate only the injected energy without the average in-time pile-up component.

Figure 5 shows the energy resolution represented by the standard deviation of $E_{\mathrm{T}}^{\mathrm{pred}}$ - $E_{\mathrm{T}}^{\mathrm{true}}$ as a function of $E_{\mathrm{T}}^{\mathrm{true}}$. Dense+RNN, Dense, and CNN architectures provide better energy resolution compared to RNN and OF, which have similar resolutions. The improvement is visible across all bins of $E_{\mathrm{T}}^{\mathrm{true}}$.

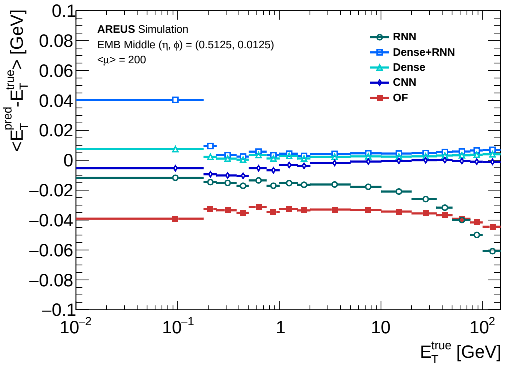

Figure 4: Transverse energy scale as a function of the true transverse energy deposited in a single cell of the electromagnetic barrel calorimeter, with a pileup of $\langle\mu\rangle = 200$.

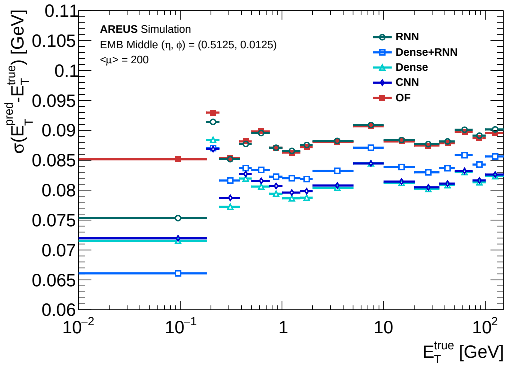

Figure 5: Transverse energy resolution as a function of the true transverse energy deposited in a single cell of the electromagnetic barrel calorimeter, with a pileup of $\langle\mu\rangle = 200$.

D. Uncertainty prediction using neural networks

The NNs shown in section C provide good energy resolution averaged over the entire sample. However, for individual energy predictions, the network has no inherent way to assess the confidence in its prediction. The deep evidential regression (DER) technique [4] enables an NN to predict uncertainties associated with its outputs. Instead of predicting a single value, the network predicts distribution parameters that describe both the prediction (mean of the distribution) and the associated uncertainty (computed from the other parameters of the distribution). The model is trained using a likelihood function (to fit the distribution) and an evidence penalty (which increases uncertainties for outliers), as a loss function.

The application of DER to energy reconstruction algorithms is investigated, as it could provide valuable information on luminosity variations, bunch train structures, or triggering capabilities. Applied to the Dense architecture, this technique provides acceptable overall uncertainty predictions, as illustrated in Figure 6. This pull plot shows that the distribution of $E_{\mathrm{T}}^{\mathrm{pred}}$ - $E_{\mathrm{T}}^{\mathrm{true}}$ divided by the predicted uncertainty (δ^{pred}) is centred near 0 with a standard deviation of 0.75 pointing to

PoS(EPS-HEP2025)078

the uncertainties being overestimated by approximately 25%. This NN maintains similar energy resolution performance to the optimised Dense model.

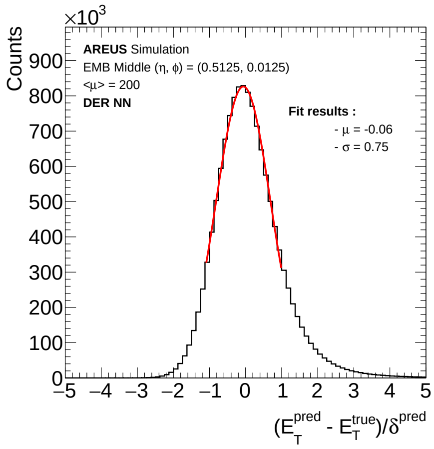

Figure 6: Pull distribution illustrated by the difference between the predicted and true transverse energies deposited in a single cell of the electromagnetic barrel calorimeter, divided by the predicted uncertainty (δ^{pred}) computed by the DER NN.

E. Conclusion

Several NN architectures have been developped to compute the energy deposited in the ATLAS LAr calorimeter on FPGA-based electronics boards. The NN hyperparameters were optimised using Bayesian optimisation to achieve the best energy resolution while minimising the model size for FPGA implementation. Three NN architectures out of the four tested outperform the legacy OF algorithm while remaining small enough to fit on FPGAs. The DER technique, applied to the Dense architecture, enables the prediction of both the deposited energies and their associated uncertainties with reasonable accuracy. The corresponding NN is still small enough to fit on FPGAs while providing a method to assess per event uncertainties.

References

[1] Georges et al. Aad. Firmware implementation of a recurrent neural network for the computation of the energy deposited in the liquid argon calorimeter of the ATLAS experiment. *JINST*, 18(05):P05017, 2023.

[2] Georges Aad et al. Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters. *Comput. Softw. Big Sci.*, 5(1):19, 2021.

[3] Georges Aad et al. Optimised neural networks for online processing of ATLAS calorimeter data on FPGAs, 2025. arXiv:2510.11469 [physics.ins-det].

[4] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1251:14927–14937, 2020.