Machine Learning on FPGAs for Real-Time Processing for the ATLAS Liquid Argon Calorimeter

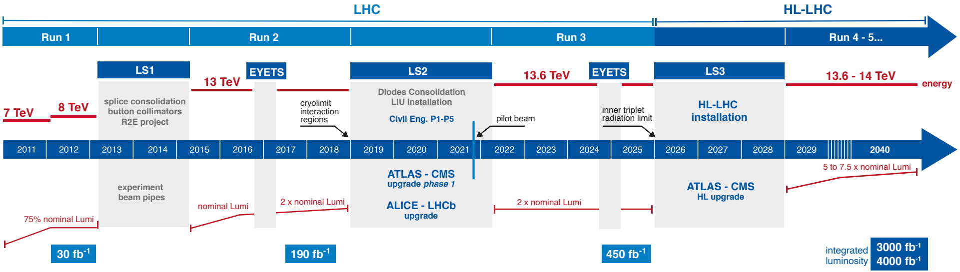
Lauri Laatu on behalf of the ATLAS Liquid Argon Calorimeter Group

20.01.2023

Content

The Phase-II Upgrade of the LHC

Upgrade of the ATLAS experiment

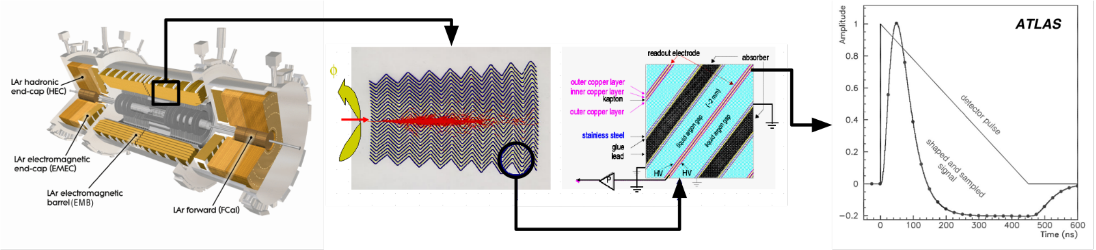


- The High Luminosity LHC (HL-LHC) is an important milestone for particle physics
  - To increase the luminosity to study rare processes
  - To increase the collision rate to up to 200 simultaneous p-p collisions (pileup) per bunch crossing
- The detectors will be upgraded to cope with the high collision rate at the HL-LHC
  - In particular the ATLAS calorimeter readout electronics will be completely replaced

ATLAS Liquid Argon Calorimeter

Energy reconstruction in the LAr calorimeter

- The Liquid Argon Calorimeter (LAr) mainly measures the energy deposited by electromagnetically interacting particles
  - Consisting of ≈ 182 000 calorimeter cells
- Passing particles ionize the material
  - Bipolar pulse shape with total length of up to 750 ns (30 BCs)
  - Pulse is sampled and digitized at 40MHz
- Energy reconstruction is done in real-time and used in triggering decision
  - Using the digitized samples from the pulse



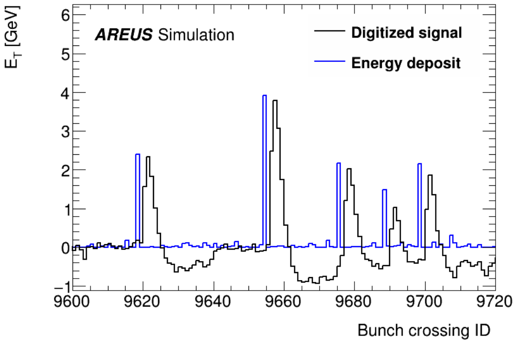
Energy Reconstruction

Energy reconstruction in the LAr calorimeter

- Current energy reconstruction uses the Optimal Filtering Algorithm with maximum finder (OFMax)

$$E(t) = \sum_{i=1}^{5} a_i \cdot s_i$$

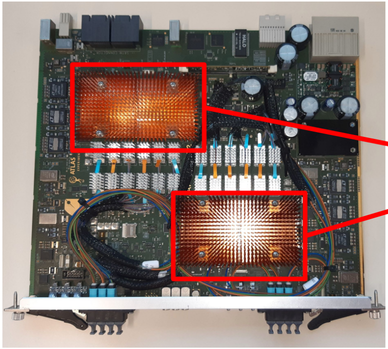
  - a_i - Predefined coefficients to fit the pulse
  - s_i - Sampled signal
- Distorted pulses result in significantly decreased performance of OFMax

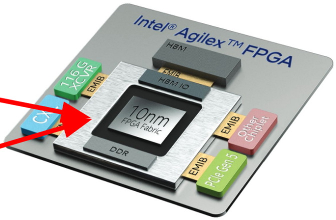


LAr Electronics Upgrade

Energy Reconstruction in Run-4

- LAr Signal Processor (LASP) board
  - For Phase-II one FPGA processes 384 channels and latency requirement of 125 ns
- Phase-II electronics with high-end FPGAs
  - Increased computing capacity
  - Improved online energy reconstruction using machine learning-based methods





The board is being tested with Intel Stratix 10 FPGAs but will be upgraded to Agilex

Table of Contents

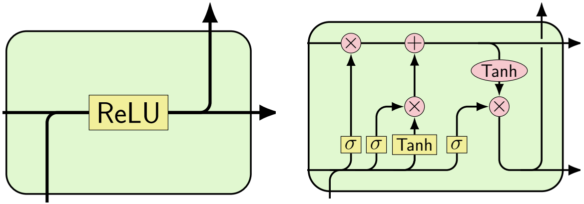
RNN Architecture

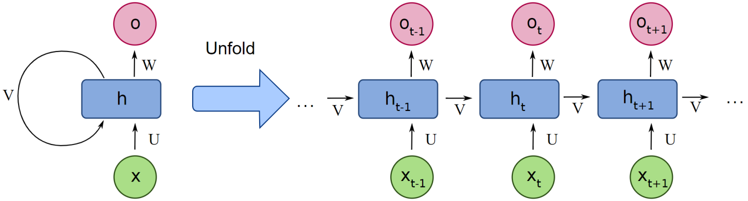
Time series processing with Recurrent Neural Networks (RNNs)

- Recurrent Neural Networks (RNNs) are designed to process time series data
- RNNs consist of neural network layers that process by combining new time input with past processed state
- Vanilla RNN is the smallest RNN structure
- Long Short-Term Memory (LSTM) network for efficiently handling past information

Vanilla

LSTM





RNNs for Energy Reconstruction

Using many-to-one and many-to-many networks for energy reconstruction

- Use digitized samples as inputs for the recurrent network
- Sliding window
  - Full sequence split into overlapping subsequences with a sliding window
  - One energy prediction per subsequence
  - Four samples in the peak, one in the past
  - Possible for Vanilla RNN and LSTM
- Single cell
  - Use the LSTM cell to process all digitized samples in one continuous chain instead of a sliding window
  - Full history of events available
  - Possible only for LSTM

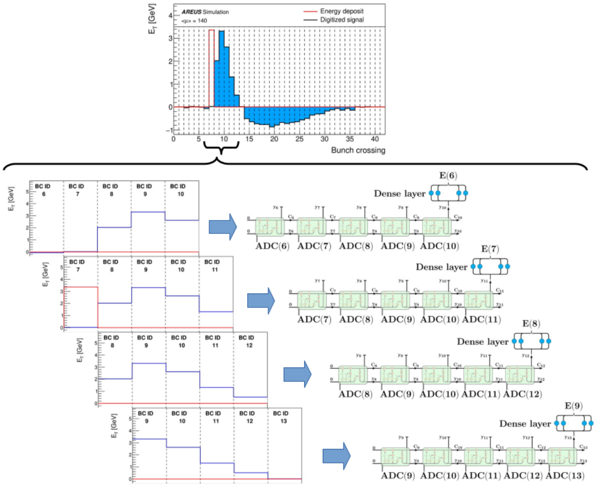
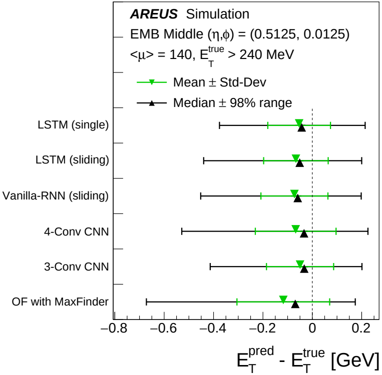


Table of Contents

NN Performance

Resolution and network size

- Overall better energy resolution than OFMax
  - Smaller tails and mean closer to zero
- Best performance with LSTM
  - Too large to fit on the FPGA
- CNNs and Vanilla RNN perform well with fewer parameters

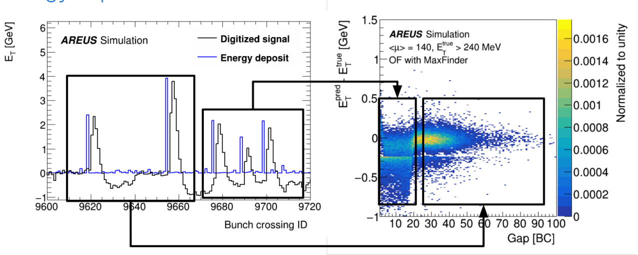


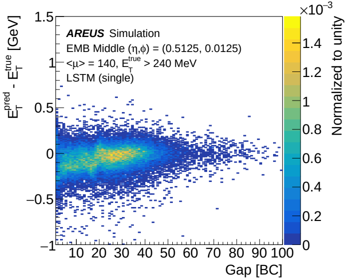
Algorithm	LSTM (single)	LSTM (sliding)	Vanilla (sliding)	CNN (3-conv)	CNN (4-conv)	Optimal filtering
Number of parameters	491	491	89	94	88	5
MAC units	480	2360	368	87	78	5

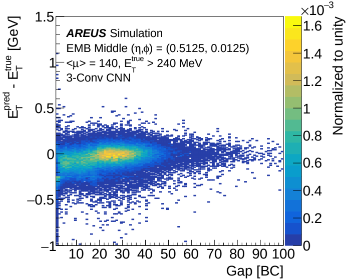
NN Performance

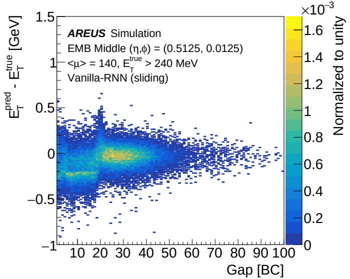
Resolution as a function of gap to previous energy deposit in BCs

- Clear performance decrease with OFMax at low gap
- All NNs perform better with overlapping events
  - More past samples allows for better correction of overlapping events









Quantization Aware Training

Optimizing NNs for firmware

- Math operations in firmware are done using fixed-point arithmetic
- Quantizing NNs after training known as post-training quantization (PTQ) with decreases the accuracy
- It is possible to mitigate this effect with quantization aware training (QAT)
    - Training using math operations as if they were quantized
- Simulation results from High Level Synthesis (HLS) implementation of RNNs show that the required bitwidth can be halved by using QAT

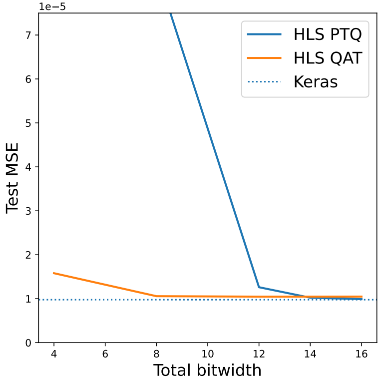
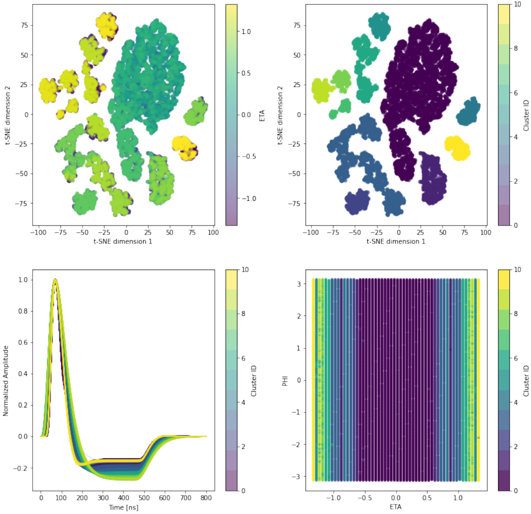


Table of Contents

Reconstruction for Full Detector

Pulse Clustering

- Pulse shape differs in the detector
  - Reduced performance with differing pulse shapes
  - One NN training will not perform well for the full detector, nor is 182k NNs feasible
  - Need to reduce the number of NNs trained while maintaining accuracy
- Clustering method used to group detector regions
  - t-SNE from calibration pulses to acquire clustering
  - DBSCAN to automatically classify cluster
  - Separation correlates with η according to pulse shape differences



Pulse Clustering

Reconstruction in different regions

- Evaluate inside same cluster
  - Train with one cell, test with another
  - Same performance as with training and testing with the same cell
- Large performance drop when training with one cluster and testing with another
- Train with mixed data from all clusters, test with single cluster
  - Mixing data across clusters slightly restores performance

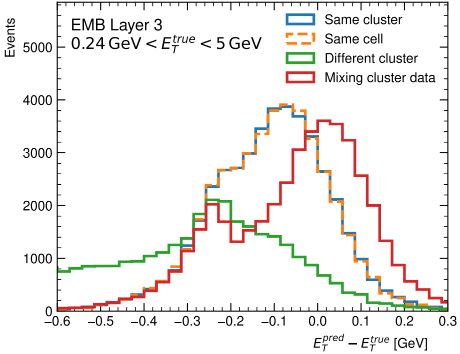


Table of Contents

Conclusion

Energy reconstruction using recurrent neural networks

- Energy reconstruction with RNNs overperforms legacy algorithms in Phase-II conditions
  - Better energy resolution overall
  - Better recovery of energy resolution with overlapping signals
- Implemented and validated in firmware and the implementations mostly fulfill the LAr real-time processing requirements
  - Testing on DevKits started and is showing good results
- Next step is to quantify the effect on object (electrons, photons) reconstruction and physics performance
- Paper published available Here