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

Lauri Laatu supervised by Emmanuel Monnier and Georges Aad

03.10.2023

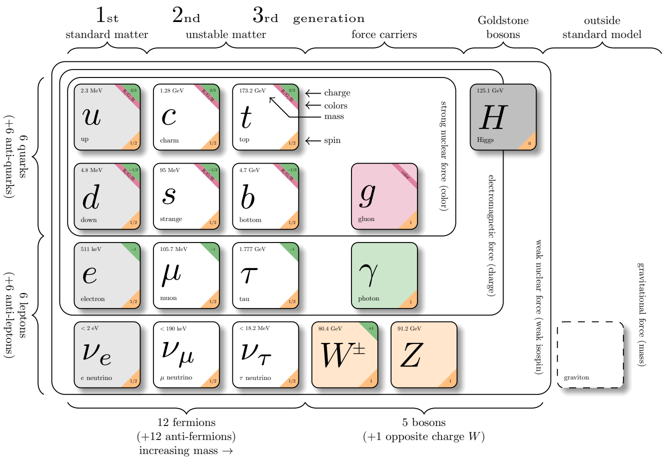
Content

1. Experimental Context
2. Network Architectures
3. Network Performance on a Single Calorimeter Cell
4. Energy Reconstruction in the Full LAr Calorimeter
5. Adapting RNNs for FPGA Deployment
6. Conclusion

The Standard Model of Particle Physics

The fundamental model of physics

- The Standard Model describes the most basic building blocks of the universe
- The Higgs Boson was found in LHC
  - No discoveries of fundamental particles since
- Current and future LHC prospects
  - Precision measurements and search of deviations from the Standard Model
  - Special focus on the Higgs sector
  - Search for Beyond Standard Model particles

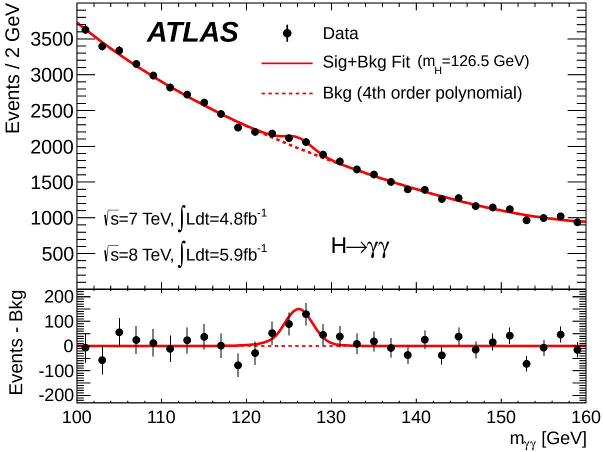


The Standard Model of Particle Physics

The fundamental model of physics

- Direct observation of resonances: fundamental way of finding new particles
  - Eg. Higgs Boson discovery in $H \to \gamma\gamma$ channel
- Fundamental role of energy resolution of the Liquid Argon calorimeter

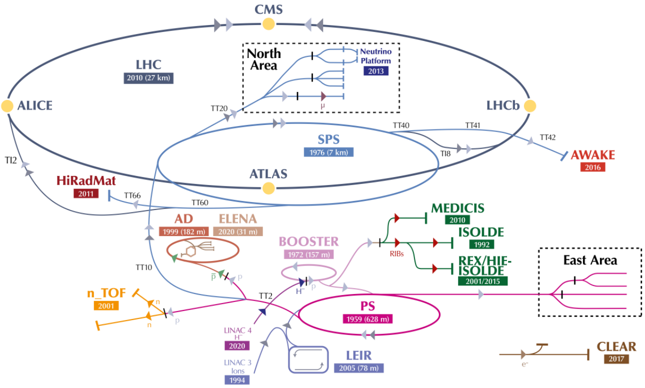
Discovery plot 4.7.2012



The Large Hadron Collider

The accelerator complex

- A 27 km circumference accelerator ring with two counter-rotating beams
- Up to 14 TeV collisions of bunches of 10^{11} protons every 25 ns (40 MHz) at interaction points

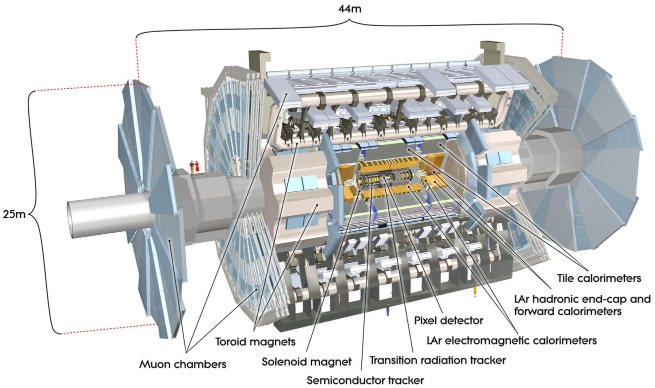


S

ATLAS

The detector

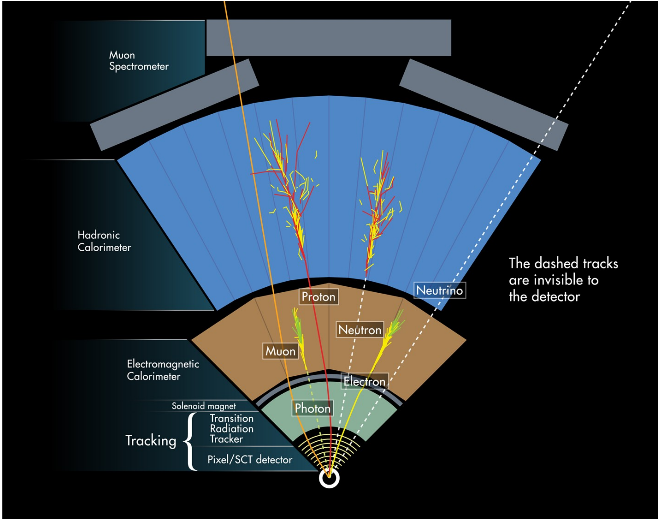
- ATLAS detector is one of the general purpose detectors in the LHC
- Each subdetector is specialized in measuring different properties of particles



ATLAS

The detector

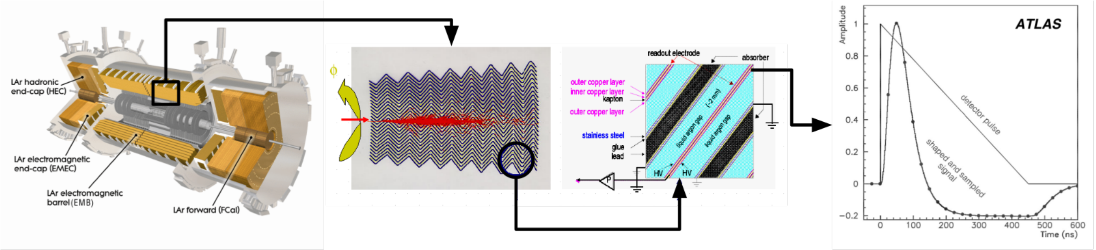
- Tracking system
  - Measures the direction and momenta of charged particles
- Electromagnetic calorimeter
  - Measures the energy of electromagnetically interacting particles
- Hadronic calorimeter
  - Measures the energy of hadronic particles
- Muon system
  - Measures the momenta of muons



ATLAS Liquid Argon Calorimeter

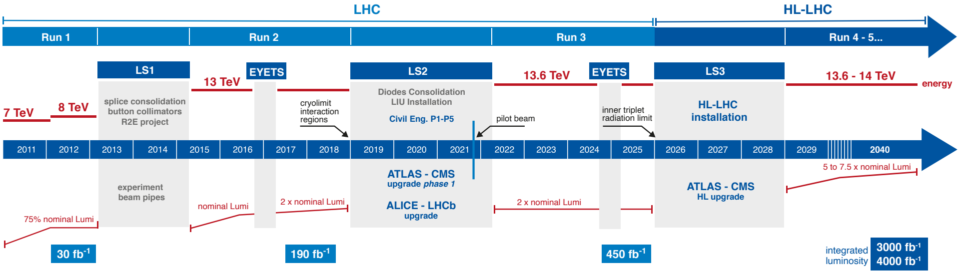
Energy reconstruction in the LAr calorimeter

- 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 800 ns (32 BCs)
  - Pulse is sampled and digitized at 40MHz
- Energy reconstruction is done using the digitized samples from the pulse
  - Computed real-time and used in triggering decision



The Phase-II Upgrade of the LHC

Upgrade of the ATLAS experiment

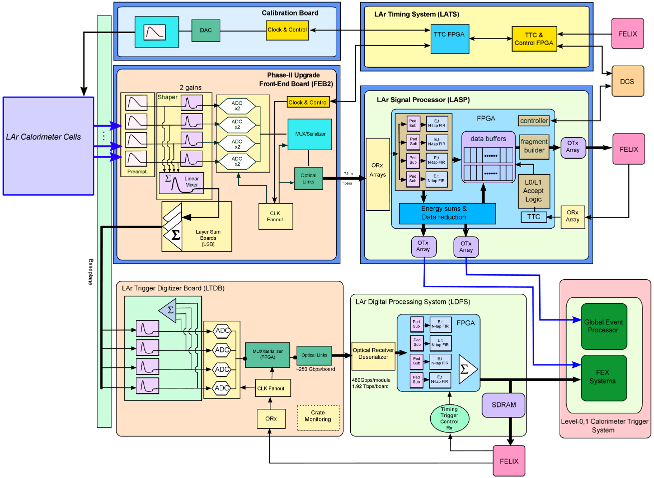


- The High Luminosity LHC (HL-LHC) is an important milestone for particle physics
  - Increase the luminosity to study rare processes
  - High pileup: up to 200 p-p collisions per bunch crossing
- The detectors will be upgraded to cope with the high pileup at the HL-LHC
  - In particular the ATLAS calorimeter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

Phase-II Upgrade of the Liquid Argon Calorimeter

LAr Electronics upgrade

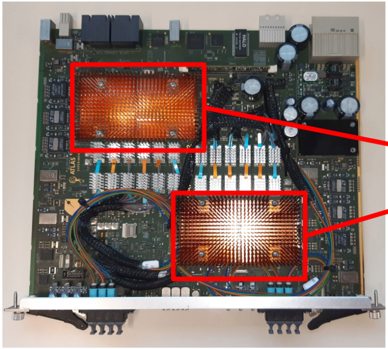
- Increased luminosity requires
  - Higher granularity at trigger
  - Increased trigger rate
- Frontend electronics amplify, shape and digitize the LAr electrical signal at 40 MHz
- Backend processing board: LASP
  - 40 MHz energy computation at full granularity
  - Send data to trigger at 40 MHz
  - Readout at 1 MHz (trigger accept)
  - Designed at CPPM

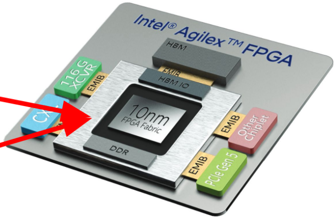


LASP Board

LAr Electronics Upgrade

- LASP board contains two high-end FPGAs (Field Programmable Gate Arrays)
  - Only technology available to handle high throughput (1Tb/s per board)
- Able to use neural networks but with FPGA constraints
  - 384 LAr channels per FPGA: small NNs
  - 125 ns latency requirement at 40 MHz: fast NNs



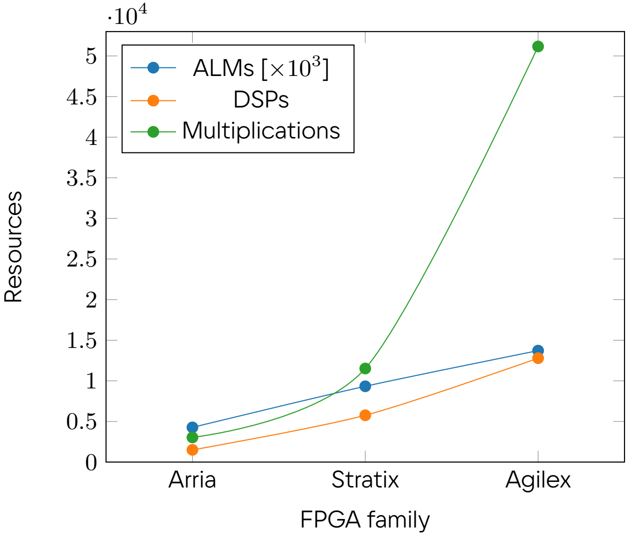


The board is being tested with Intel Stratix 10 FPGAs but will be upgraded to Agilex

FPGA Resources

Comparison of FPGAs

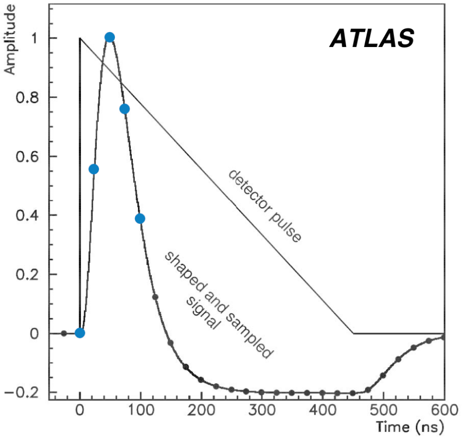
- Arria is used in Phase-I boards
- Stratix in the demonstrator board
- Agilex chosen for the production boards
- Important resources for NNs
  - Adaptive Logic Modules (ALMs) for additions
  - Digital Signal Processor (DSPs) for multiplications
- Arria and Stratix allows two multiplications with a single DSP
- Agilex allows four multiplications per DSP



Energy Reconstruction

Energy reconstruction in the LAr calorimeter

- Current energy reconstruction uses the Optimal Filtering algorithm (OFMax)
- Energy reconstruction from amplitude
- Timing τ from pulse phase changes
  - Quality and bunch crossing identification
  - Searches of long-lived particles



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

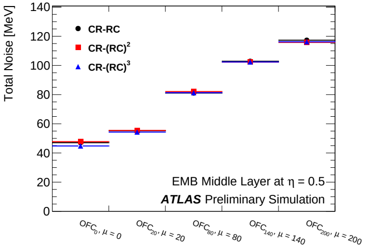
- a_i - Coefficients for energy reconstruction
- s_i - Sampled signal

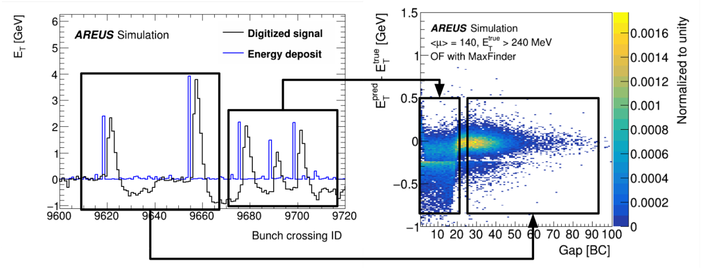
$$\tau(t) = \frac{1}{E} \cdot \sum_{i=1}^{5} b_i \cdot s_i$$

- b_i - Coefficients for timing
- s_i - Sampled signal

OF Performance in Phase-II

Increasing pileup





- Reduced performance as average pileup $\langle \mu \rangle$ increases
- Large drop in performance in the case of overlapping pulses (low time gap)
- Requires a better energy reconstruction method: **neural networks**

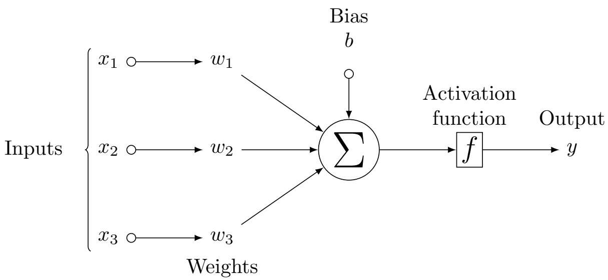
Table of Contents

1. Experimental Context
2. Network Architectures
3. Network Performance on a Single Calorimeter Cell
4. Energy Reconstruction in the Full LAr Calorimeter
5. Adapting RNNs for FPGA Deployment
6. Conclusion

Neural Networks

The neuron as the building block

- Neuron is the building block in neural networks
- Activation: non-linear function allowing learning non-linear relationships in data

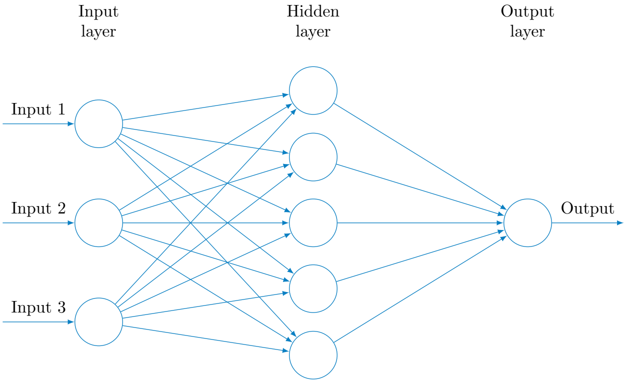


$$y = f_a\left(\sum_{i=1}^{3} w_i \cdot x_i + b\right)$$

Neural Networks

Feed Forward Networks

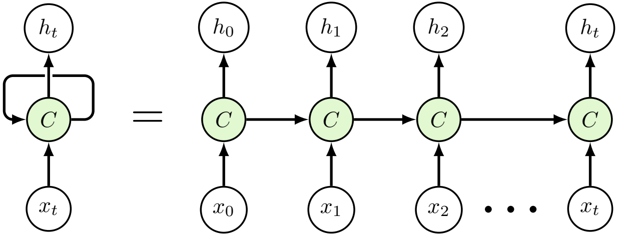
- Neural networks consist of stacked layers of neurons
- Known as the Feed Forward Network or **Dense** network



RNN Architecture

Time series processing with Recurrent Neural Networks (RNNs)

- Recurrent Neural Networks (RNNs) are designed to process time series data
  - Natural choice for processing ADC sequence from the LAr detector
- The RNN cell combines new time input with past processed state
  - Information about the past encoded in the state
  - Suitable to correct pileup from past events

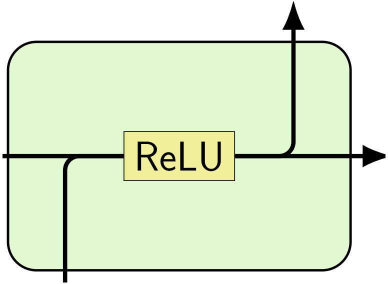


RNN Cell

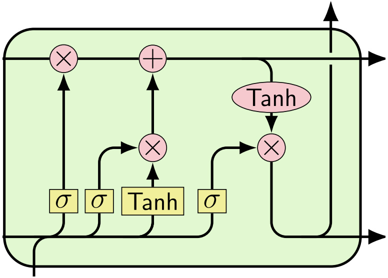
RNN cell types

- Simple RNN is the smallest RNN structure
  - Fewer parameters, limited memory
  - Rectified Linear Unit (ReLU) activation
- Long Short-Term Memory (LSTM) network for efficiently handling past information
  - More parameters, gated structure improves memory

Simple RNN



LSTM



RNNs for Energy Reconstruction

Using sliding window RNN for energy reconstruction

- Full sequence split into overlapping subsequences with a sliding window
- One energy prediction per subsequence
  - 4 samples after the energy deposit
  - 1 before the deposit (in this example)
- Longer sequence adds samples in the past
  - Better correction of out-of-time pileup
- Single neuron reconstructs energy from the RNN state

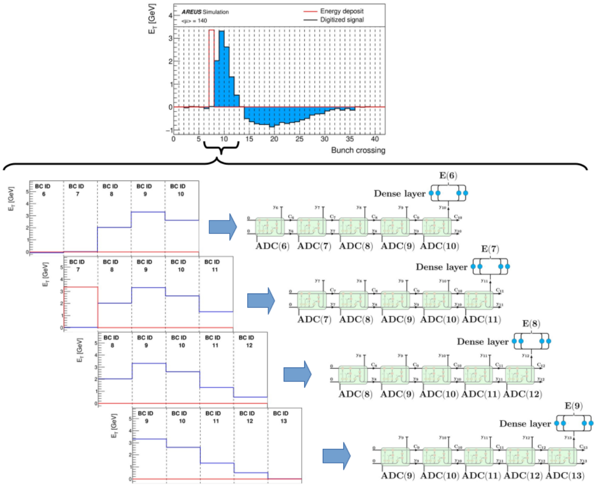


Table of Contents

1. Experimental Context

2. Network Architectures

3. Network Performance on a Single Calorimeter Cell

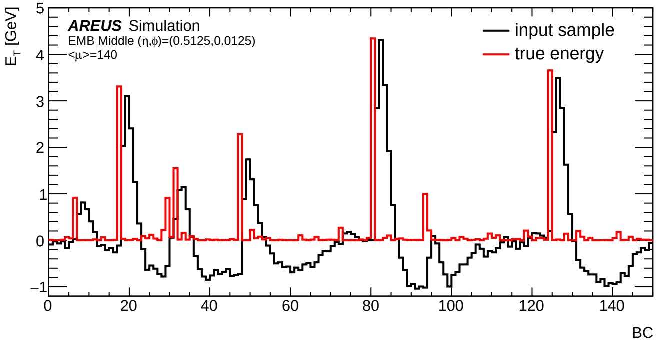
4. Energy Reconstruction in the Full LAr Calorimeter

5. Adapting RNNs for FPGA Deployment

6. Conclusion

ATLAS Simulated Dataset

Simulation of a single calorimeter cell



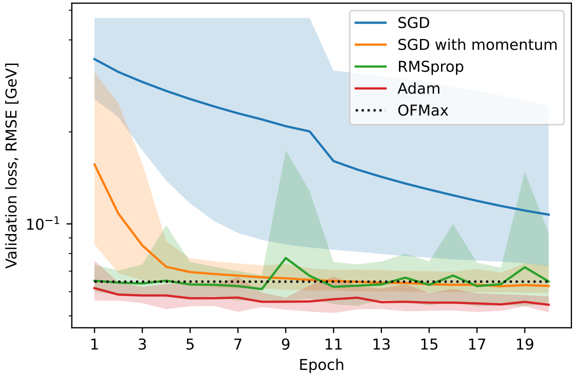
- EMB Middle
  - $\eta = 0.0125$
  - $\phi = 0.5125$
- Pileup $\langle\mu\rangle = 140$
- Injected signal
  - $E_T \in [0 \text{ GeV}, 5 \text{ GeV}]$
  - Gap: $\overline{X} = 30$, $\sigma = 10$

- ATLAS Readout Simulator (AREUS) is used for continuous simulation of a single LAr cell
  - Minimum bias events at different $\langle\mu\rangle$ levels superimposed with injected pulses

Training Simple RNN for Energy Reconstruction

Hyper-parameter optimization

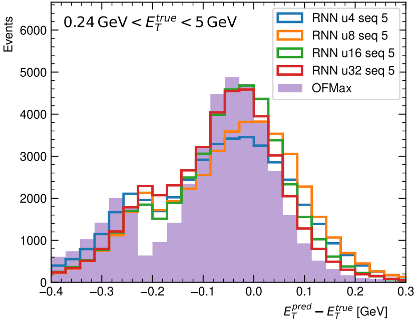
- RNNs developed in Keras
- Optimization for different parameters
    - Optimizer: minimization algorithm
    - Loss: minimization function
    - Epoch: number of iterations over the dataset
    - Batch size: training samples per optimization step
    - Choice of activation function for Simple RNN and LSTM

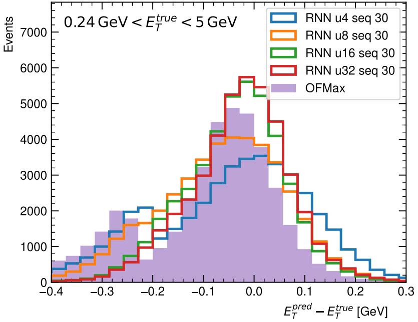


Optimizing Simple RNN for Energy Reconstruction

Simple RNN performance

- Out-of-time pileup causes large drop in performance for OF
- RNNs with sufficient size and sequence length correct for out-of-time pileup

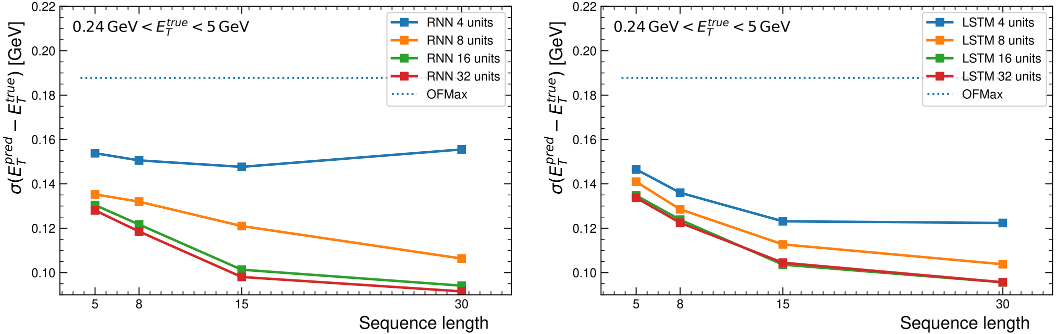




Optimizing Simple RNN and LSTM for Energy Reconstruction

Summarizing performance

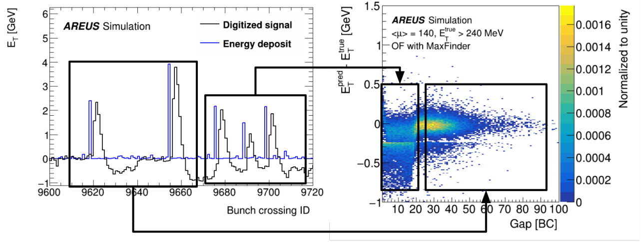
- Optimization for Simple RNN and LSTM architectures
- Performance in $\sigma(E_T^{pred} - E_T^{true})$ as function of sequence length
- Requires enough units and long sequence



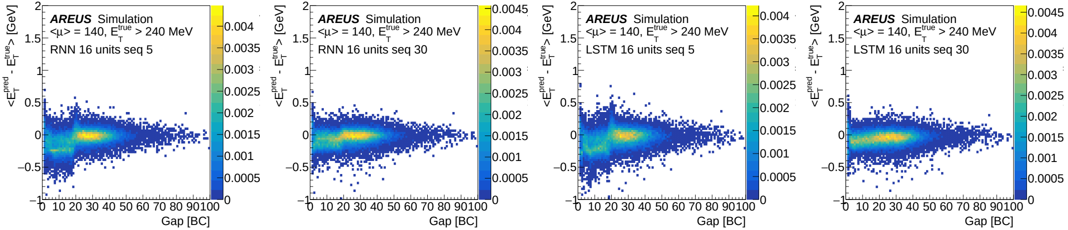
RNN Performance

Resolution as a function of gap to previous energy deposit in bunch crossings (BCs)

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



Simple RNN **LSTM**

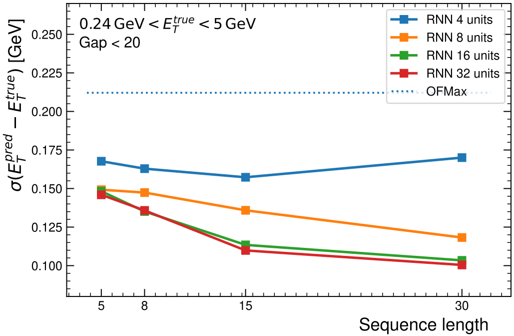


Simple RNN Gap Performance

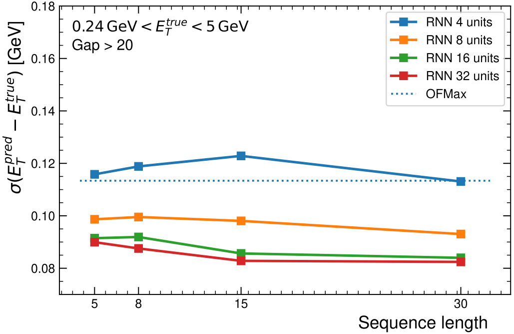
Simple RNN performance in low and high gap region

- Separation in overlapping and non-overlapping cases
- Overall improvement with more units
- Longer sequence length improves performance in the case of overlapping pulses

Overlapping



Non-overlapping

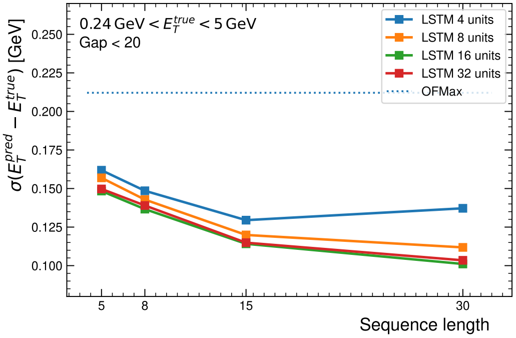


LSTM Gap Performance

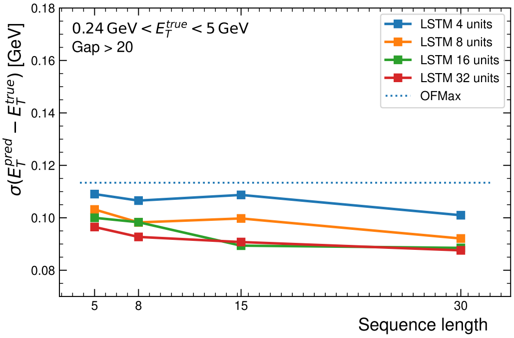
LSTM performance in low and high gap region

- Separation in overlapping and non-overlapping cases
- Overall improvement with more units
- Longer sequence length improves performance in the case of overlapping pulses

Overlapping



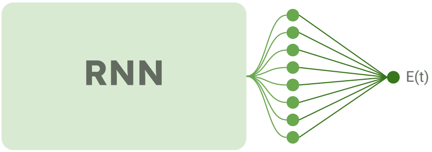
Non-overlapping

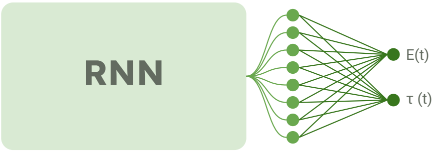


Timing Reconstruction

RNNs predict time and energy

- Extending the Dense layer to include another output neuron for the timing
- The same RNN state is reconstructed as both the energy (E^T_{pred}) and timing (T_{pred} or τ)

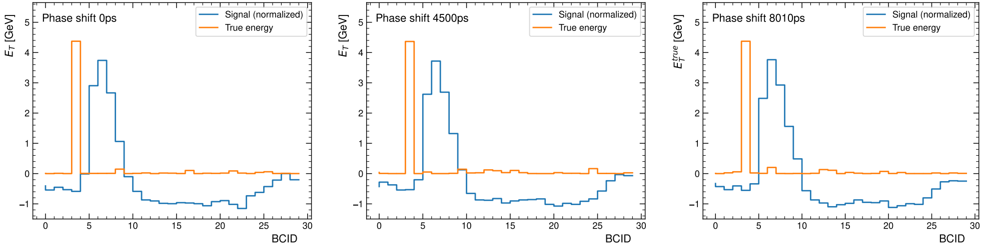




Timing Reconstruction

Dataset for timing predictions

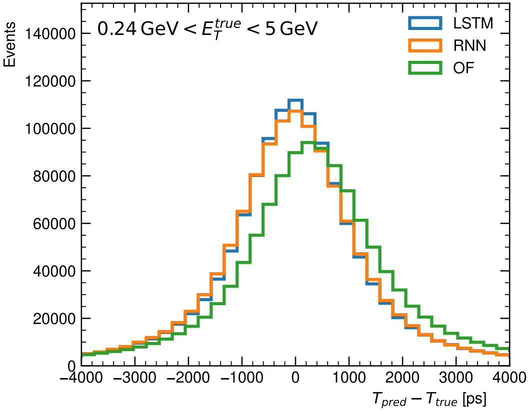
- Changes in the time of energy deposit alters the pulse shape
- Dataset consist of uniform -8 ns to 8 ns shift



Timing Predictions

Timing prediction performance

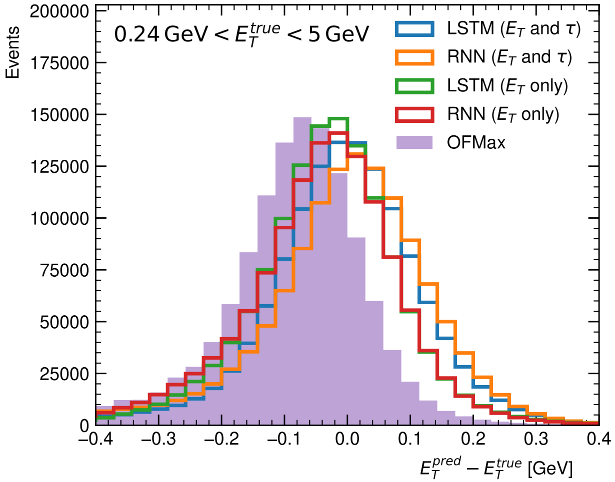
- RNNs outperform OF in timing reconstruction
  - σ(LSTM) = 1.82 ns, σ(RNN) = 1.83 ns, σ(OF) = 2.87 ns



Timing Predictions

Timing and energy prediction performance

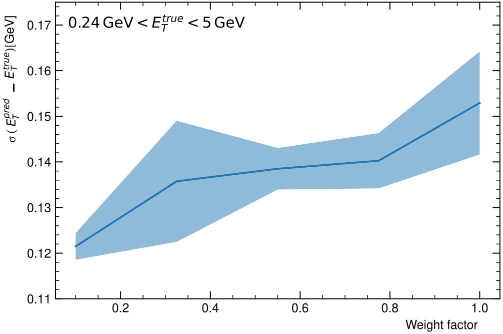
- Noticeable drop in energy reconstruction performance
- The training is prioritizing timing resolution at the expense of energy resolution

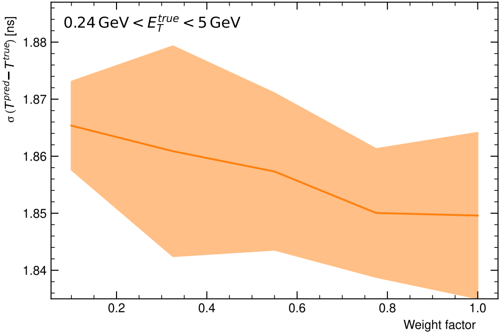


Weighted Loss

RNNs predict time and energy

- Poor energy resolution can be mitigated by implementing a custom loss function with a parameter to prioritize energy resolution more
- Loss function incorporates a weight factor to prioritize energy at the expense of timing reconstruction performance: $MSE(E^T_{pred} - E^T_{true}) + w \cdot MSE(T_{pred} - T_{true})$

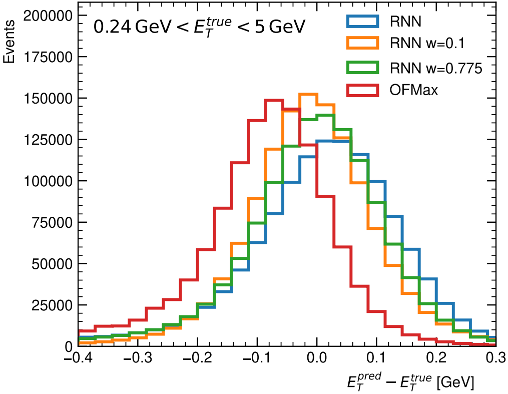


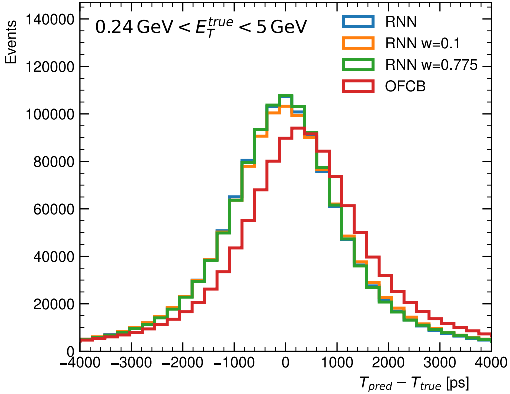


RNNs with Weighted Loss

RNNs predict time and energy

- RNNs with optimized loss outperform OF in both timing and energy resolution

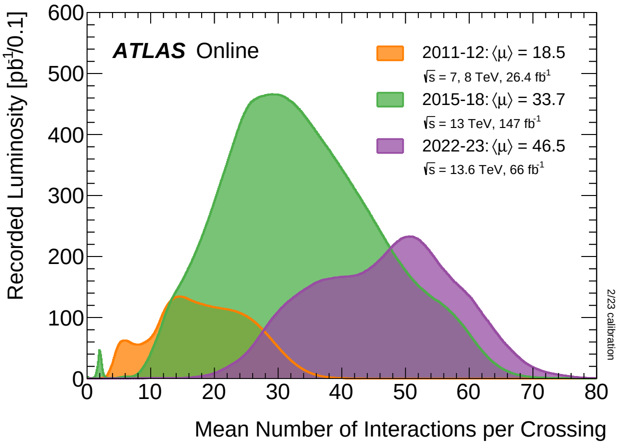




Resilience Against Instantaneous Luminosity

ATLAS luminosity across runs

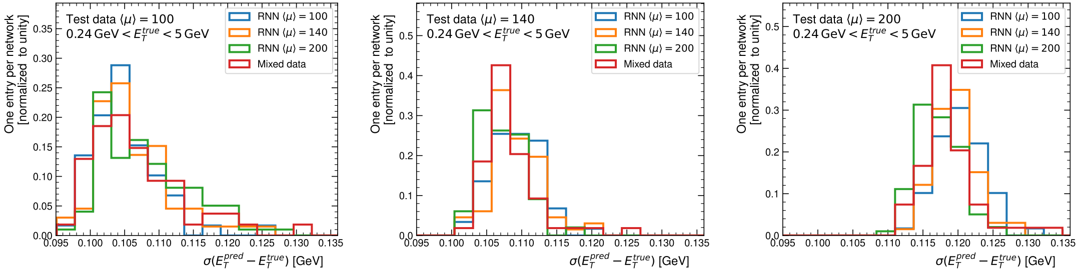
- The pileup rate changes
  - Between runs
  - During a run
- RNNs need to be resilient against the changes in the luminosity



Resilience Against Instantaneous Luminosity

RNN performance over a range of luminosities

- Train 80 randomly initialized networks for $\langle\mu\rangle = 100$, $\langle\mu\rangle = 140$, $\langle\mu\rangle = 200$ and by mixing $\langle\mu\rangle$ data
    - Probe statistical uncertainty of network training
- The effect of $\langle\mu\rangle$ is smaller than the effect of initialization



Resilience Against Instantaneous Luminosity

Best performing RNN networks over a range of luminosities

- Best performing RNN for each $\langle\mu\rangle$ evaluated in comparison to OFMax
- The RNNs are robust and outperform OF with different instantaneous pileup

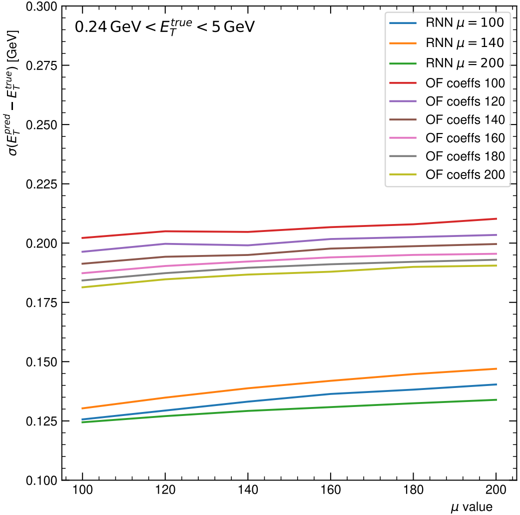


Table of Contents

1. Experimental Context

2. Network Architectures

3. Network Performance on a Single Calorimeter Cell

4. Energy Reconstruction in the Full LAr Calorimeter

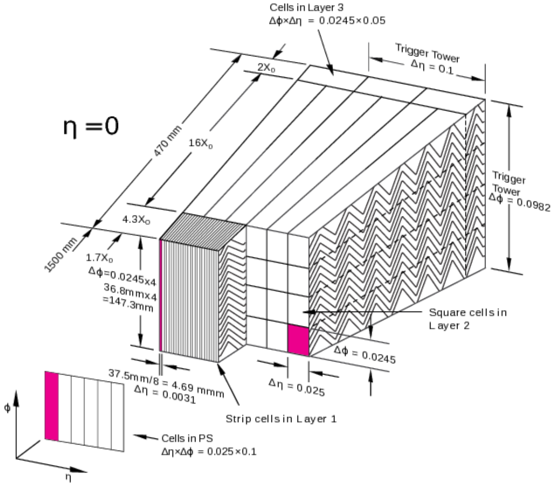
5. Adapting RNNs for FPGA Deployment

6. Conclusion

Changes in Detector Conditions

Cell clustering background

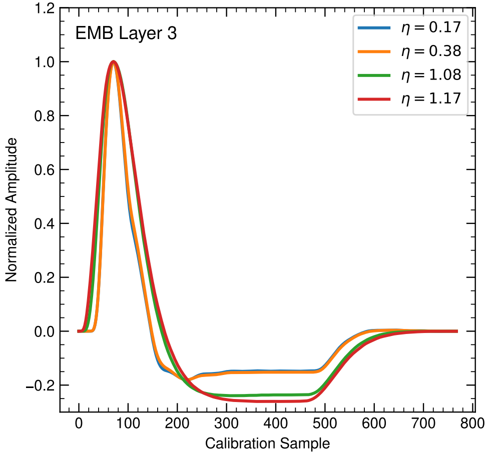
- Unfeasible to train 182k networks
- Differences of detector response in different parts in the detector
  - Leads to changes in pulse shapes
- Requires a way to identify detector cells that could share the same RNN network



Reconstruction for Full Detector

Unsupervised learning for calibration pulses

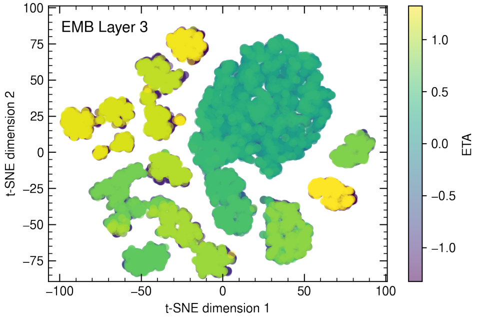
- Regular calibration process used to acquire the pulse shape of each 182k detector cells
- These calibration pulses can be used by unsupervised learning algorithms group cells together based on their similarity



t-SNE Dimensionality Reduction

Finding similarities in calibration pulses

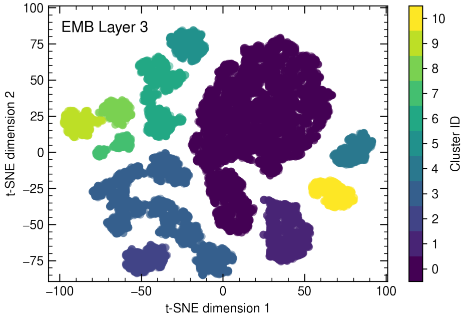
- t-SNE unsupervised learning method used for dimensionality reduction
- The method reduces the 768 samples of the calibration to two dimensions
  - Points in 2D attract or repel each other based on their similarity in higher dimension
- Similar shapes are grouped together
  - Color denotes η
  - Acquired the expected grouping based on η



DBSCAN for Cluster Labeling

Automatically detect the amount of clusters

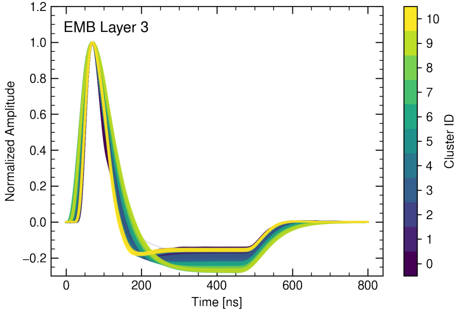
- DBSCAN is used for identifying the clusters
  - Based on their distance in the two dimensional plane
- The number of clusters is automatically detected
- Results to labeled clusters

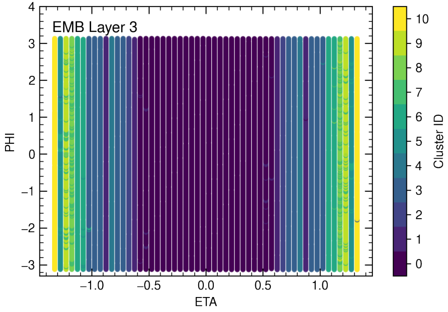


Clustering Result

Pulse shapes in the clusters

- Pulse shapes in cluster ID color
  - Similarly shaped pulses clustered together
- Calorimeter cells in $\eta - \phi$ plane shown in cluster colors
  - Expected η dependence

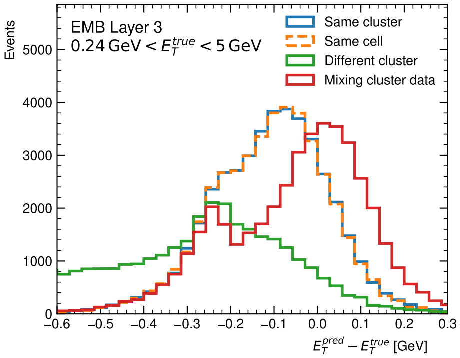




Pulse Clustering

Reconstruction across calorimeter cells in different clusters

- 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



Reduction in the Amount of Networks

Clustering for EMB and EMEC

- Clustering was done for full electromagnetic calorimeter
  - Yields a significant reduction in required amount of NNs
- In total 121 networks needed for EMB and EMEC
  - More clusters in EMEC due to large differences in geometry

Layer	Barrel	End-cap
0	7732 $\rightarrow$ 9	1521 $\rightarrow$ 3
1	58172 $\rightarrow$ 2	28259 $\rightarrow$ 26
2	28893 $\rightarrow$ 6	23185 $\rightarrow$ 46
3	13682 $\rightarrow$ 11	10138 $\rightarrow$ 18

Table of Contents

1. Experimental Context

2. Network Architectures

3. Network Performance on a Single Calorimeter Cell

4. Energy Reconstruction in the Full LAr Calorimeter

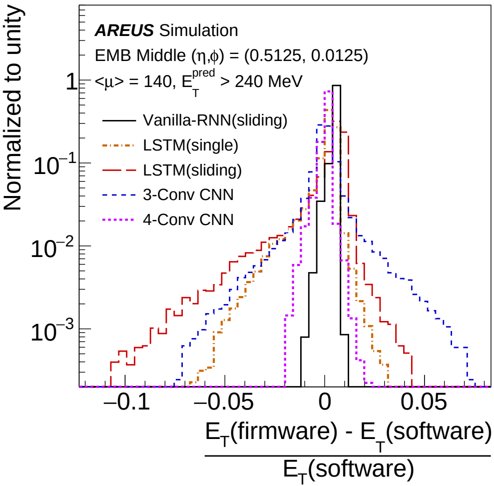
5. Adapting RNNs for FPGA Deployment

6. Conclusion

Quantization of RNNs

Deploying NNs on FPGAs

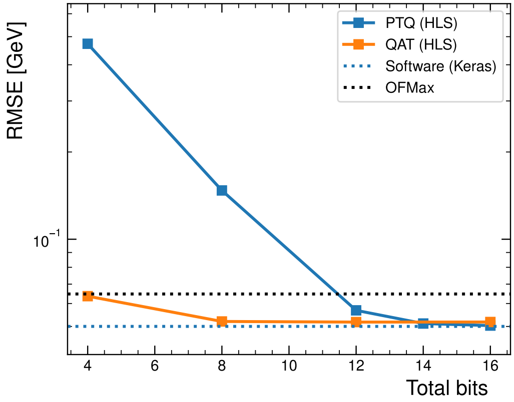
- Floating point operations on FPGAs consume large amount of resources
- Requires the usage of fixed-point operations
  - Introduces quantization error
- Optimizing the number of bits to reduce error
  - Using 18 bit precision
  - Quantization done after training
  - Shows good agreement between the software and firmware results
- Resource usage can be reduces with lower bitwidth



Quantization Aware Training

Fixed-point math operations

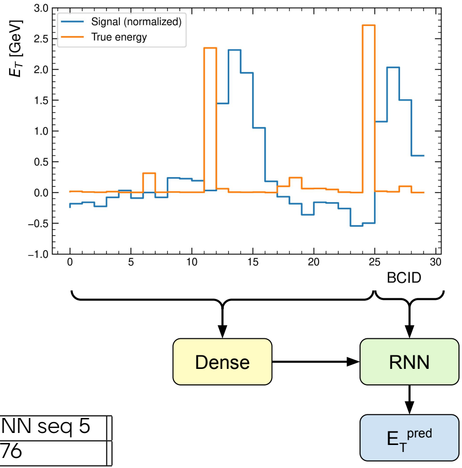
- Quantization Aware Training (QAT) uses fixed-point precision during training
- Post-Training Quantization (PTQ) requires 14 bits to reach software precision
- QAT reaches same precision with only 8 bits
- Lower resource usage per operation allows the deployment of larger network
- Agilex FPGAs have 9x9 DSP mode which can be utilized only with QAT



Resource efficient correction

Correct for past events by setting the initial state of RNN

- Long sequence length improves performance
  - Each sample adds another RNN cell iteration
- Out-of-time pileup corrected by a Dense layer
  - 25 past samples processed by the dense layer
  - Output of the dense is the initial state of the RNN



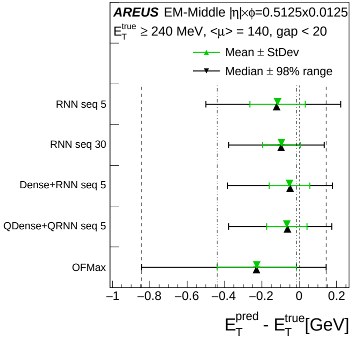
Model	RNN seq 5	RNN seq 30	Dense+RNN seq 5
Multiplications	1376	8176	1776

Resource efficient correction

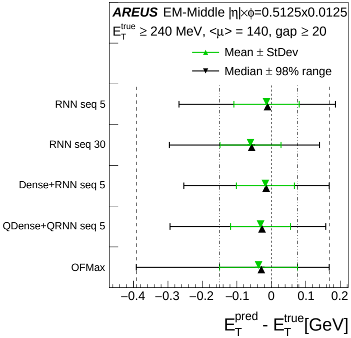
Performance of the new simple RNN architecture using 16 units

- RNN seq 5
  - 1 sample before deposit
- RNN seq 30
  - 26 samples before deposit
- Dense+RNN seq 5
  - The new architecture
- QDense+QRNN seq 5
  - 9 bit QAT

Overlapping



Non-overlapping



Similar performance with reduced resource usage

Table of Contents

1. Experimental Context
2. Network Architectures
3. Network Performance on a Single Calorimeter Cell
4. Energy Reconstruction in the Full LAr Calorimeter
5. Adapting RNNs for FPGA Deployment
6. Conclusion

Conclusion

Energy reconstruction using recurrent neural networks

- HL-LHC conditions require improved method for energy computation
- Developed RNNs outperform OF in both energy and timing resolution
- RNNs shown to be robust against changes in luminosity
- Unsupervised learning used to reduce the required amount of NNs
  - 4 orders of magnitude reduction in required networks
- Optimized for deployment on FPGA
  - Quantization Aware Training makes 9x9 bit DSP mode available
  - Out-of-time pileup corrections with Dense layer
- Two papers published
  - Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters (https://doi.org/10.1007/s41781-021-00066-y)
  - Firmware implementation of a recurrent neural network for the computation of the energy deposited in the liquid argon calorimeter of the ATLAS experiment. (arXiv:2302.07555)
- Conference talk: VIII international conference on High Energy Physics in the LHC Era (HEP2023)

Future Prospects

Energy reconstruction using recurrent neural networks

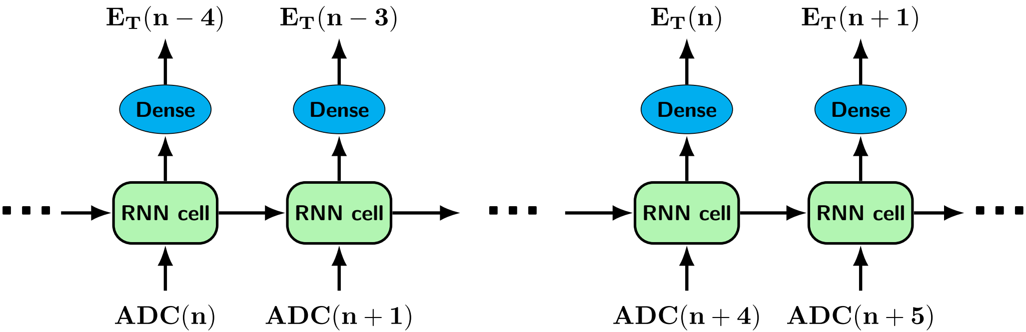
- Further tuning and improvement of the neural network
  - Explore other network architectures (transformers)
- Further optimization for FPGAs
  - Pruning of RNN: removal of insignificant weights
- Evaluate the improvement of RNNs on physics objects
  - Photon/electron identification and energy resolution
  - Trigger efficiency
- Application of neural networks on real data
  - Calibrate the network performance in real data
  - Estimate the systematic uncertainty

Backup

RNNs for Energy Reconstruction

Single-cell

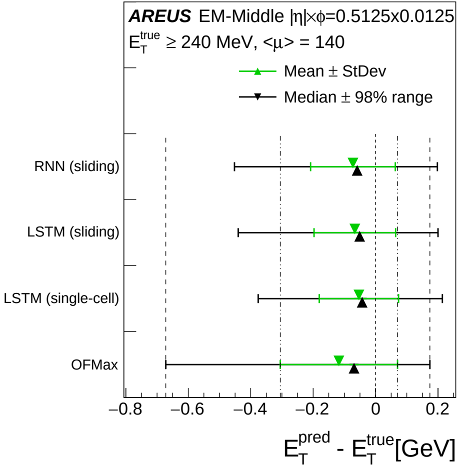
- Using LSTM to predict a continuous stream of digitized samples
  - Use the LSTM cell to process all digitized samples in one continuous chain instead of a sliding window
  - Full history of events available
  - Possible only for LSTM



Single-cell LSTM for Energy Reconstruction

Continuous stream

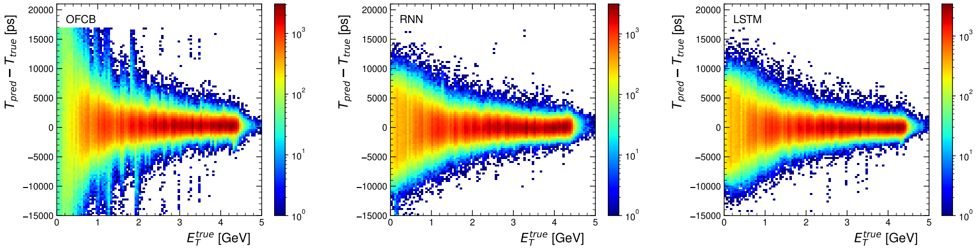
- Well performing architecture
- However unfeasible
  - Unable to implement to ATLAS simulation software
  - Unable to implement on FPGA



Timing Reconstruction

RNNs predict time and energy

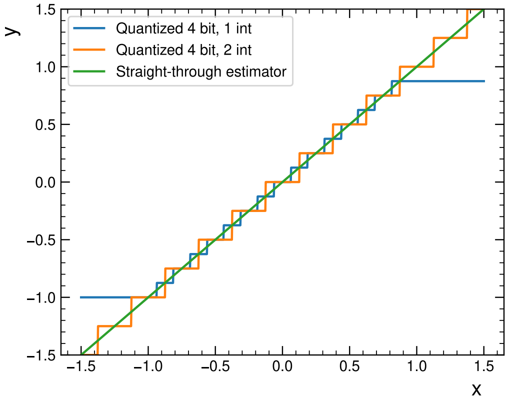
- Good timing reconstruction performance in comparison to OF
- Large improvement in lower energies



Quantization Aware Training

Fixed-poing math operations

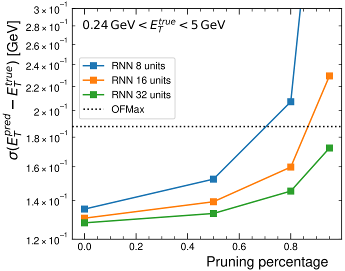
- Train neural networks with fixed-point representation
  - Simulated quantized representation in the forward-pass
  - Floating-point during weight adjustment

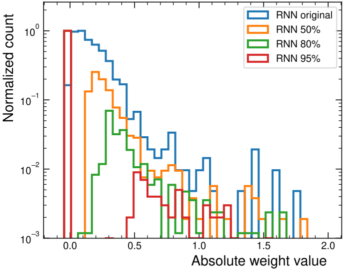


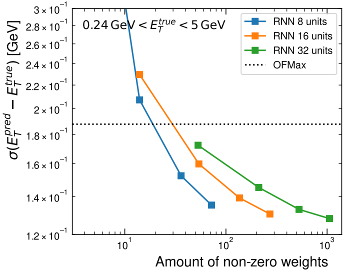
Pruning of RNNs

Removing insignificant weights

- Pruning refers to the removal of insignificant weights
- Deployment on FPGAs gives the option to omit the weight multiplication if the weight is zero
- Comparison of different pruning percentages on different sized networks
  - Larger network keeps higher performance with high pruning percentage
  - Unpruned small network still performs better than pruned large network
  - Unfeasible hardware implementation







Clustered Performance

Performance in the full EMB layer

- Evaluating performance over all of the clusters in the layer
- Train a network for each cluster by using the pulse shape in the middle of the cluster as the pulse shape for the AREUS simulation
- Evaluate the network for 10 randomly sampled cells of every cluster