25th IEEE Real Time Conference - Elba, Italy

A 1D-Convolutional Autoencoder for Pulse Compression in Nuclear DAQ Systems

J.M. Deltoro[1,2], V. González[1], A. Gadea[2], G. Jaworski[3], A. Goasduff[34]
[1]University of Valencia, Spain; [2]Instituto de Física Corpuscular, CSIC-UV, Spain; [3]Heavy Ion Laboratory, University of Warsaw, Poland; [4]Laboratori Nazionali di Legnaro, Italy

Abstract

Modern nuclear physics experiments, such as the NEDA detector, face severe data bandwidth bottlenecks in their data acquisition systems. To address this challenge, we present a real-time edge computing compression architecture using a **Quantized Split 1D-Convolutional Autoencoder deployed on an FPGA**. The system was trained using **Quantization-Aware Training (QAT)** and physically evaluated through a Hardware-in-the-Loop (HIL) setup using a **PYNQ-Z2** board. Our results demonstrate that it is possible to achieve data compression (**32:1**) directly on the edge hardware with minimal latency and a highly efficient logic footprint. Crucially, despite the hardware optimizations and fixed-point arithmetic, the system preserves signal integrity, maintaining the accurate Neutron-Gamma discrimination capabilities required by the physics of the experiment.

Introduction

NEDA is crucial for reaction channel selection in nuclear structure experiments. However, its high-frequency sampling (200 MHz) generates a data deluge that exceeds current DAQ bandwidth capacities, especially during high-count rate experiments. Standard transmission of raw 256-sample traces saturates communication channels, forcing a reduction in beam intensity or an increase in costly facility time.

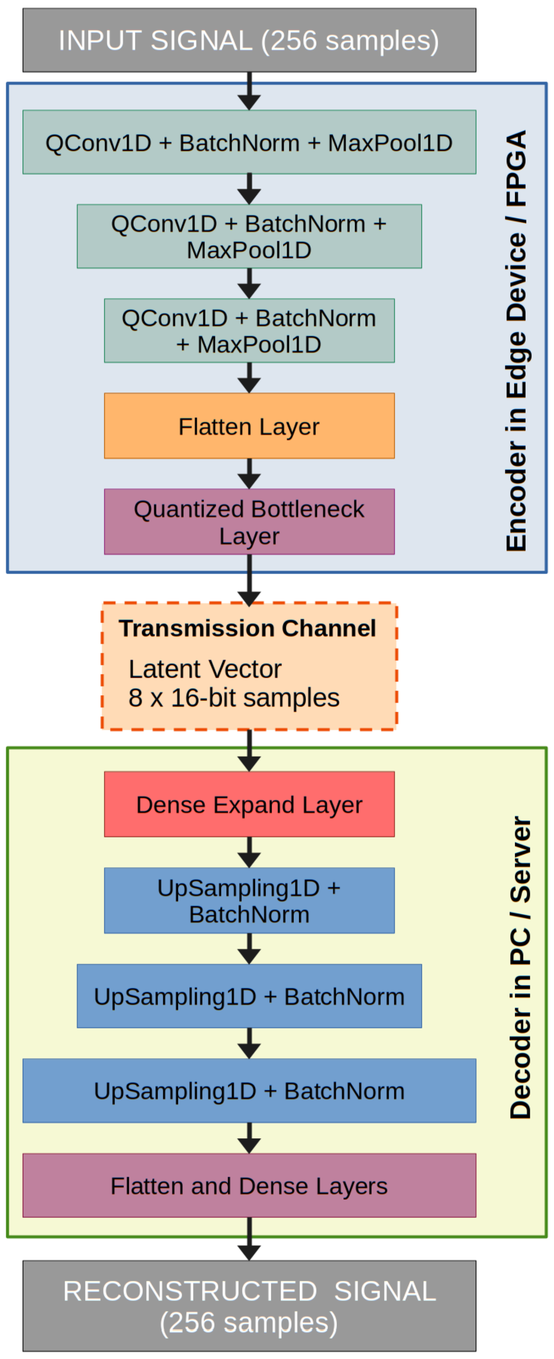

Figure 1. Schematic representation of the proposed split 1D-CAE architecture

Split 1D-Convolutional Autoencoder

To overcome this, we propose a Split 1D-Convolutional Autoencoder (1D-CAE) architecture:

- **Edge Compression:** A quantized encoder (QAT) resides in the FPGA, reducing each pulse from 256 16-bit samples to only 8 features of 16-bit integers (32:1 compression ratio).
- **Cloud Reconstruction:** A decoder on the server-side reconstructs the pulses for offline analysis.
- **Goal:** Maximize event throughput and minimize energy consumption without sacrificing the physical integrity required for particle identification.

Software Validation (QKeras)

Before physical deployment, the model was validated in software using **QAT to simulate fixed-point arithmetic**.

- **Physics Preservation:** The **Charge Comparison (CC)** distributions for Neutron/Gamma discrimination show negligible degradation compared to uncompressed ADC data (**Global Accuracy of 95.71%**).
- **Signal Fidelity:** Reconstructed traces maintain an excellent **Mean Correlation** with the original of **0.992**.

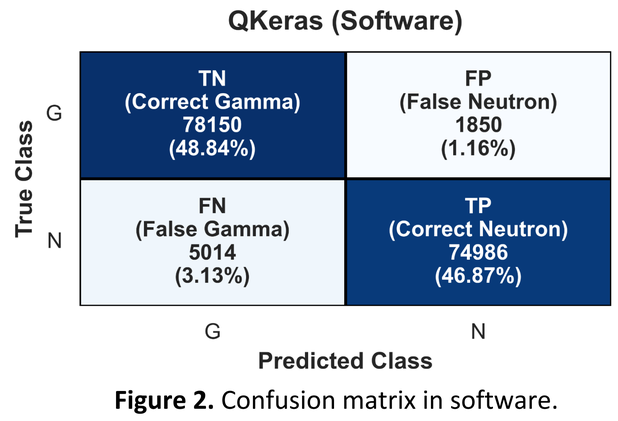

Figure 2. Confusion matrix in software.

Hardware Edge Implementation

The encoder was synthesized using **HLS4ML** and validated using a HIL TCP setup on a **PYNQ-Z2 board**.

- **Physics Preservation and Signal Fidelity:** Global Accuracy of **92.56%**. Mean Correlation with the original **0.989**.
- **Real-Time Processing:** Deterministic latency of **74.7 µs** per event @100MHz.

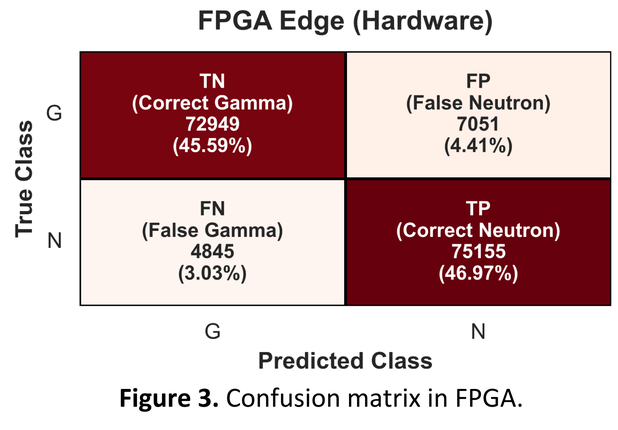

Figure 3. Confusion matrix in FPGA.

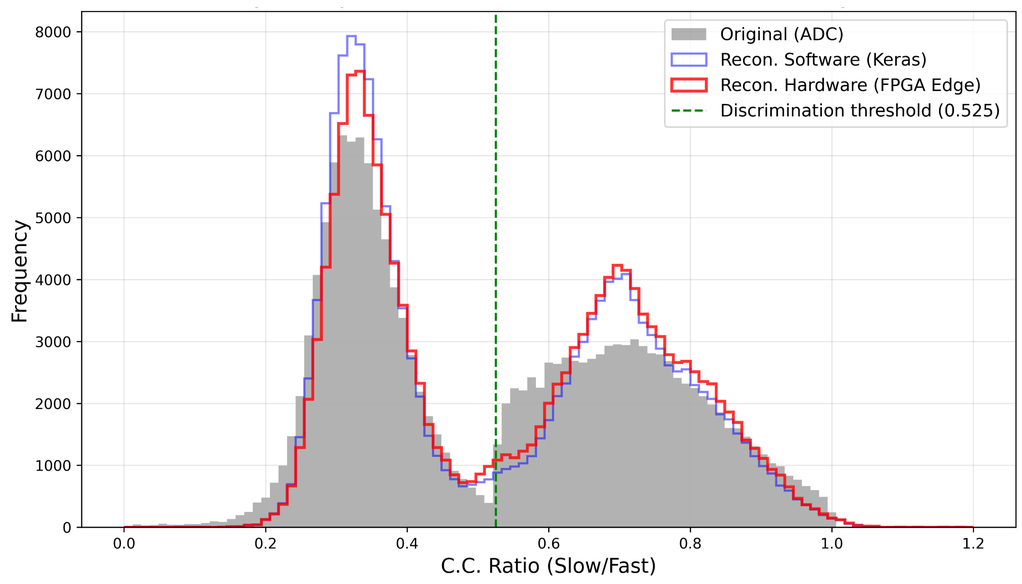

Figure 4. Comparison of CC ratios obtained with original signals, software compression, and with FPGA compression. 80,000 events were taken for each particle type (neutron and gamma).

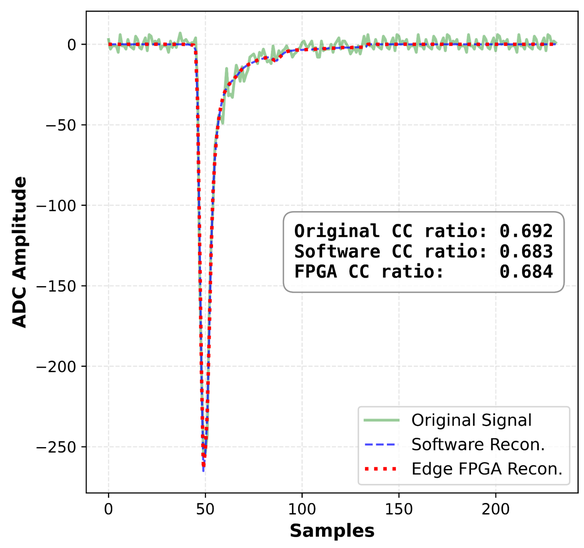

Figure 5. Example of the original signal and the reconstructed signal after software compression and after FPGA compression

Conclusions

- **Bandwidth Reduction:** Achieved a **32:1 compression ratio** directly at the edge.
- **Real-Time Execution:** Validated via HIL in a PYNQ-Z2, the IP core operates with a deterministic latency of just 74.7 µs per event.
- **Physics Preservation:** Accurate Neutron-Gamma discrimination is maintained.
- **Future Work:** Full integration into the NEDA front-end DAQ firmware for on-the-fly data compression during in-beam experiments.

Contact

Jose Manuel Deltoro Berrio - Universitat de València, Valencia, Spain
Email: josemadeltoro@uv.es Phone: +34 650 11 82 76
Website: https://www.linkedin.com/in/jose-manuel-deltoro/

References

1. Deltoro, J.M., Jaworski, G., Goasduff, A. et al. Reconstruction of pile-up events using a one-dimensional convolutional autoencoder for the NEDA detector array. NUCL SCI TECH 36, 32 (2025). https://doi.org/10.1007/s41365-024-01606-y
2. Valiente-Dobón, J., Jaworski, G., Goasduff, A., et al., "NEDA—NEutron Detector Array," Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 927, pp. 81–86, May 2019. https://doi.org/10.1016/j.nima.2019.02.021
3. Duarte, J., Han, S., Harris, P. et al. Fast inference of deep neural networks in FPGAs for particle physics. Journal of Instrumentation, 13(07), P07027 (2018). https://doi.org/10.1088/1748-0221/13/07/P07027
4. Coelho, C.N., Kuusela, A., Li, S. et al. Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. Nat Mach Intell 3, 675–686 (2021). https://doi.org/10.1038/s42256-021-00356-5

Acknowledgment: We would like to acknowledge Mohamed Benseddiq for his valuable technical support in the setup and configuration of the PYNQ-Z2 system.

Funding: Work funded by MCIN/AEI/10.13039/501100011033 and Generalitat Valenciana, Spain with grants MCIU PRTR-C17.I01 and Generalitat Valenciana (GVANEXT) ASFAE/2022/031 and by the EU NextGenerationEU funds.