FPGA-Deployed Variational Autoencoder for Real-Time Soft X-Ray Electron Temperature Reconstruction in RFX-mod2

L. Orlandi[1,2] A. Rigoni Garola[2] L. Saccaro[1,2] P. Franz[2] M. Gobbin[2,3] L. Piron[1,2] R. Cavazzana[2]

[1]Centro Ricerche Fusione, Università di Padova, Padova, Italy
[2]Consorzio RFX (CNR, ENEA, INFN, Università di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy
[3]Istituto per la Scienza e la Tecnologia dei Plasmi, CNR, Padova, Italy

1. Introduction

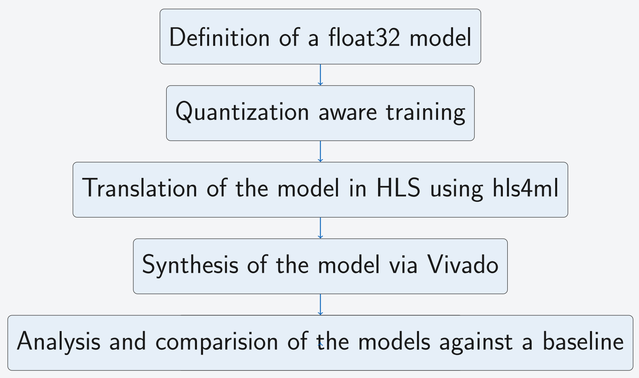
- New fusion experiments need reliable diagnostics that can supply data during D-T campaigns, given the constraints that prevent hardware adjustments for issue resolution.
- There is a critical need for a system capable of compensating for any missing data.
- Conventional forecasting systems are inadequate for real-time application.
- Artificial intelligence presents novel opportunities for data retrieval in real-time.

Research Objective

The objective of this research is to evaluate the efficacy of a quantized neural network model in providing real-time diagnostic data during operational phases, while maintaining high accuracy and achieving optimal speed necessary for integration into the control loop.

2. Methodology

Workflow Overview



Technical Details

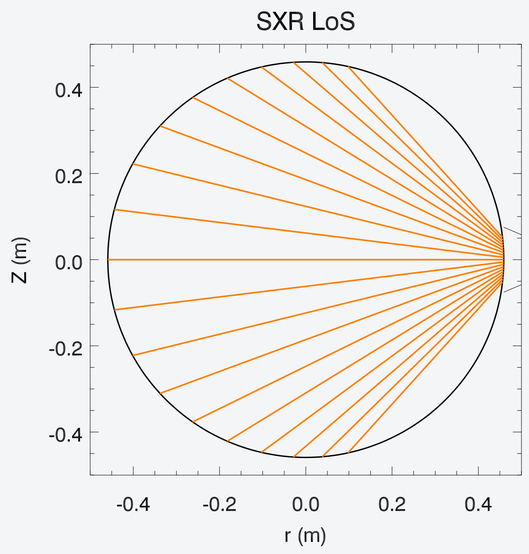
- Dataset: ∼70,000 electron temperature profiles from the DSX3 diagnostic (Soft X-Ray)
- Variational Auto-Encoder (VAE) in its beta (β) variation
- Beta annealing procedure (Warm-up and cyclical)
- Loss: MSE, AKL
- Metrics: L1 Distance on selected points to reconstruct

3. Key Contributions

- Developed a model for reconstructing 1D signals from various diagnostics.
- Computed maximum error for float32 and quantized model for real-time reconstruction on FPGA.
- Conducted latency measurements for float32 and quantized model in order to evaluate their suitability for real-time application.

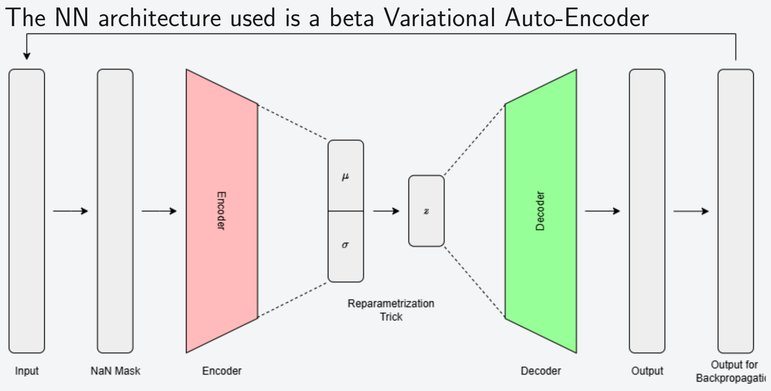
4. Diagnostic

DSX3 This diagnostic is used in combination with the two-filter technique to compute the electron temperature of the plasma starting form the plasma brightness (line integral of the plasma emissivity).



5. Neural Network Architecture

The NN architecture used is a beta Variational Auto-Encoder



This architecture was selected owing to its capability for generalizing the learned distribution, thereby enabling effective data imputation.

6. Experimental Results

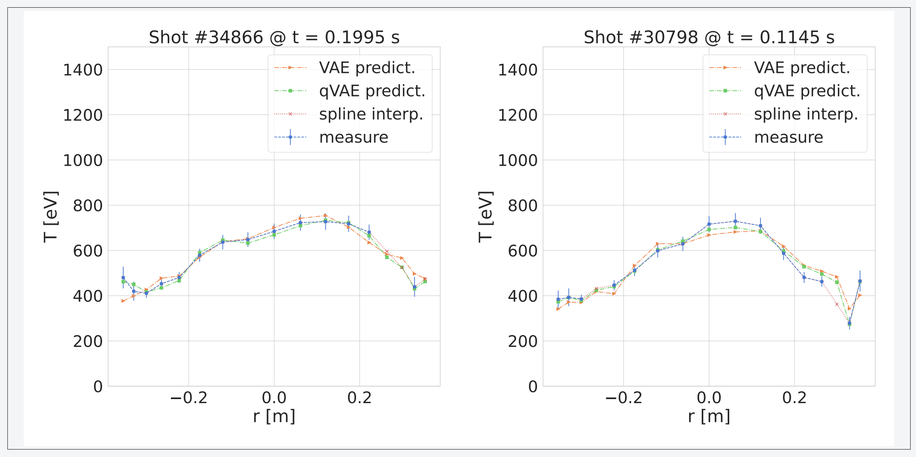


Figure 1. The developed methodologies demonstrate a high-fidelity reconstruction of electron temperature profiles. It is evident that the float32 model exhibits greater accuracy compared to its quantized counterpart, though the difference is not substantial.

Figure 2. The error comparison between the float32 and quantized models reveals that the quantized model performs marginally worse than the float32 model overall. However, there is an unexpected improvement in performance for cases with seven missing points.

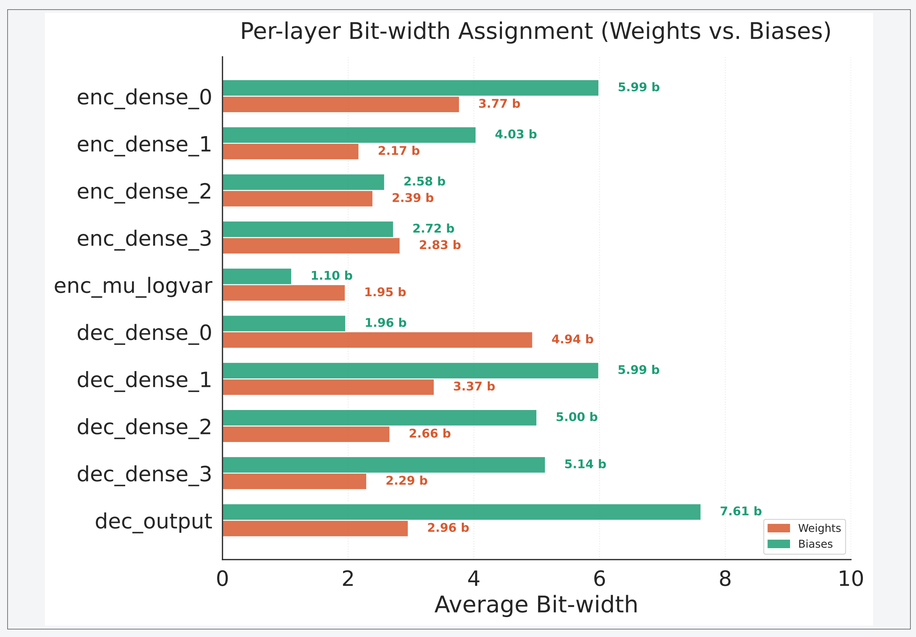


Figure 3. The results of quantization-aware training for the model are presented in the figure, which displays the average bit-widths for each layer. This visualization effectively highlights the location of the bottleneck within the network.

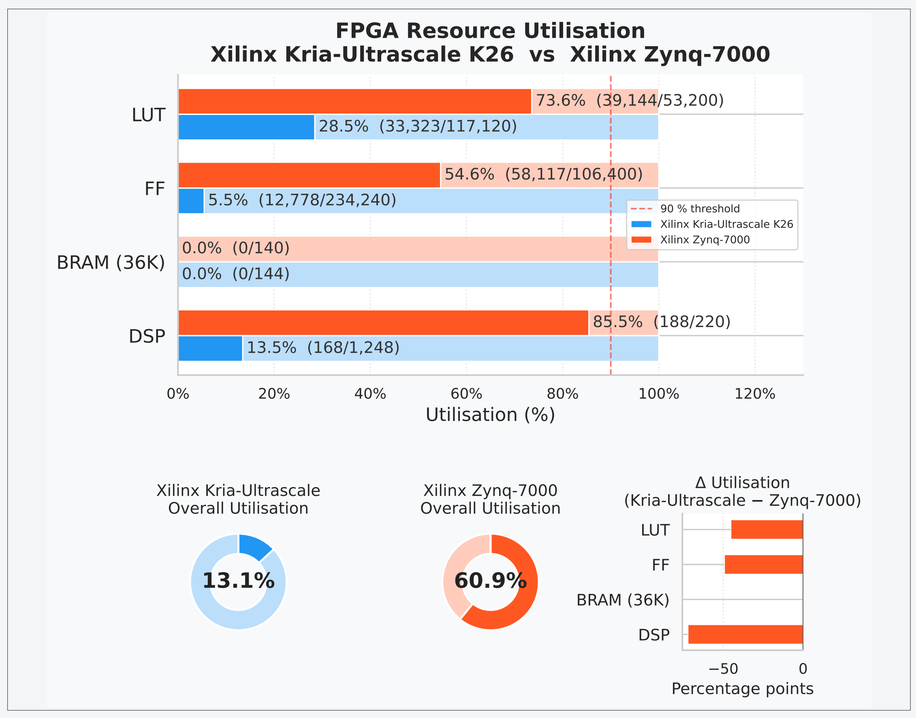


Figure 4. Resource utilization for the quantized model when synthesized for the target FPGAs (Kria and Zynq).

7. FPGA implementation

- The quantized model was synthesized using Vivado.
- Initially, the project targeted the Kria FPGA, however, it is possible to implement it also on smaller FPGAs such as the Zynq.
- The resource usage for the model on both cards is lower than what the cards allow (Figure 4).
- A comparative analysis of latency between the two investigated FPGAs and the GPU implementation of the model is provided in Table 1

Implementation	Latency [μs]	Throughput [MS/s]	EBOPs
Nvidia Tesla P40	31	0.032	$7.2 \cdot 10^6$
AMD Krya xck26	0.140	7.1	$9.6 \cdot 10^4$
AMD Zynq xc7	0.792	1.3	$9.6 \cdot 10^4$

Table: Resource consumption and latency/throughput for different cards

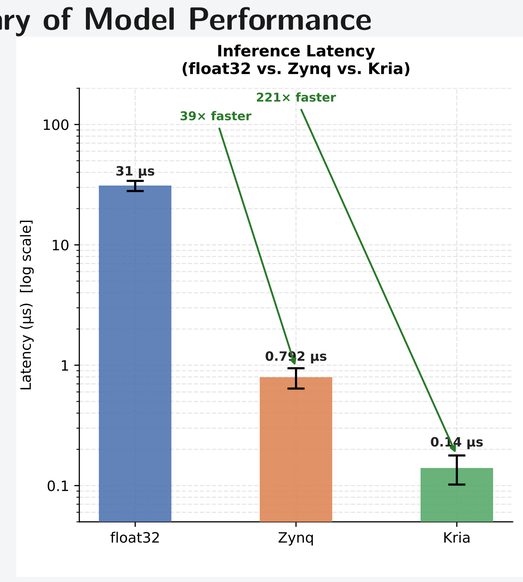
8. Discussion

- Quantization aware training slightly affects the model's ability to reconstruct electron temperature profiles.
- The quantization penalty is outweighed by the model's seamless deployment on FPGA, enabling real-time implementation in experiments.
- The translation from Python to HLS using hls4ml was successful, with identical performance across tested instances of the test dataset.
- Vivado synthesis was issue-free, confirming the models compatibility with small FPGAs like the AMD Zynq.

9. Conclusion

- The workflow used has been validated to be perfect for implementation of real-time computing on fusion devices such as RFX-mod2.
- The quantized model maintains accuracy despite noisy inputs and missing data points.
- The low resource usage allows deployment on various FPGAs and potentially mixing different models on a single board if requirements are met.

Summary of Model Performance



10. Future Work

- Real-time deployment and testing on a working device. Currently not possible because RFX-mod2 is still under upgrade.
- Integrate latent space with data from magnetic probes to enhance reconstruction accuracy.
- Parallel deployment with other neural networks encoding various diagnostics, fuse latent representations to form a concise plasma state representation.

11. References

[1] S.Peruzzo et al., "Design concepts of machine upgrades for the RFX-mod experiment", SOFT-29, 2016

[2] F.Bonomo et al., "A multichord soft x-ray diagnostic for electron temperature profile measurements", Rev. Sci. Instrum. 2006

[3] A.Rigoni Garola et al., "Diagnostic data integration using deep neural networks for real-time plasma analysis", IEEE Trans. 2021

[4] P.Franz et al., "Experimental investigation of electron temperature dynamics of helical states in the rfx-mod reverse field pinch", NF 2013

[5] L.Orlandi et al., "Data reconstruction using variational autoencoders and error analysis compared to b-spline interpolation", PPCF 2026

12. Contact Information

Email: luca.orlandi@igi.cnr.it

Email: luca.orlandi.1@phd.unipd.it

This work is co-funded by the European Union under the Next Generation EU initiative, ENI SpA and Università degli studi di Padova. This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No. 101052200 - EUROfusion). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.