

Evaluating Device Variability in RRAM-Based Single- and Multi-Layer Perceptrons

Alan Blumenstein^{1,2}, Eduardo Pérez^{3,4}, Christian Wenger^{3,4}, Nadine Dersch^{1,2}, Alexander Kloes¹,
Benjamín Iñíguez², Mike Schwarz¹

¹ NanoP, THM, Giessen, Germany

² DEEEA, Universitat Rovira i Virgili, Tarragona, Spain

³ IHP - Leibniz Institute for High Performance Microelectronics, Frankfurt (Oder), Germany

⁴ BTU Cottbus-Senftenberg, Cottbus, Germany

SUMMARY

This work evaluates the effect of stochastic weight variations in resistive random-access memory (RRAM)-based implementations of artificial neural networks (ANNs). It studies two types of ANNs: the Single-Layer Perceptron (SLP) [1] and the Multi-Layer Perceptron (MLP) [2]. The study focuses on comparing their sensitivity under two types of device-level variability: device-to-device (D2D) and cycle-to-cycle (C2C).

A simulation framework is developed using a Variable Neural Network (VNN) model [3], where Gaussian noise is applied to synaptic weights to emulate the statistical behavior of RRAM devices. The extent of variability is controlled by an Adjustment Rate (AR), which defines the proportion of perturbed weights. Quantization is also introduced, with 9 and 21 discrete levels considered, to examine the effect of resolution on accuracy under variability.

Results using the MNIST dataset show that SLPs are significantly more sensitive to variability, with accuracy dropping rapidly as AR increases. In contrast, MLPs demonstrate greater robustness and more gradual performance degradation. Increasing quantization levels from 9 to 21 consistently improves accuracy stability, especially for MLPs. Stochastic quantization further enhances performance in MLPs but has minimal impact on SLPs.

The impact of AR on accuracy is analyzed for the D2D simulation of the MLP with quantization at two levels. As shown in Fig. 1, increasing quantization from 9 to 21 levels improves median accuracy from 77.87% to 83.66%, with the 21-level setup showing greater stability. Similarly, for the SLP Fig. 2, illustrates a drop in median accuracy from 75.41% (21 levels) to 46.85% (9 levels) as AR increases, again highlighting the higher stability achieved with more quantization levels.

This approach helps identify the impact of synaptic weight variation and optimal quantization levels for hardware by analyzing their effects on ANN's accuracy (η), robustness, and AR. It enables the selection of optimized network architectures and practical quantization strategies that balance performance and resource efficiency. Understanding AR's influence supports hardware design by ensuring performance goals are met with minimal overhead, improving efficiency, robustness, and cost-effectiveness.

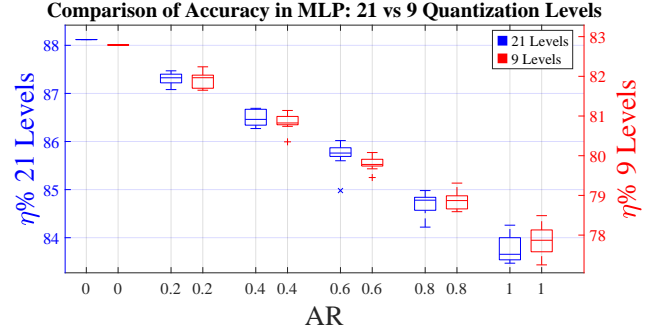


Fig. 1. MLP performance with 21 Quantization levels vs 9 Quantization levels

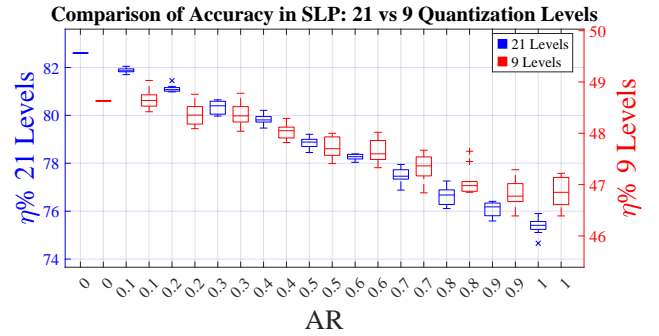


Fig. 2. SLP performance with 21 Quantization levels vs 9 Quantization levels

REFERENCES

- [1] N. Dersch, E. Perez-Bosch Quesada, E. Perez, C. Wenger, C. Roemer, M. Schwarz, and A. Kloes, "Efficient circuit simulation of a memristive crossbar array with synaptic weight variability," *Solid-State Electronics*, vol. 209, p. 108760, Nov. 2023.
- [2] V. Milo, F. Anzalone, C. Zambelli, E. Perez, M. K. Mahadevaiah, O. G. Ossorio, P. Olivo, C. Wenger, and D. Ielmini, "Optimized programming algorithms for multilevel rram in hardware neural networks," in *2021 IEEE International Reliability Physics Symposium (IRPS)*. IEEE, Mar. 2021, pp. 1–6.
- [3] A. Blumenstein, E. Pérez, C. Wenger, N. Dersch, A. Kloes, B. Iñíguez, and M. Schwarz, "Exploring variability and quantization effects in neuronal networks using the MNIST dataset," *accepted to 11th Joint EuroSOI Workshop and International Conference on Ultimate Integration on Silicon (EuroSOI-ULIS 2025)*, 2025.