Toward reliable gradient-based training on stochastic memristive devices

K. Nikiruy[(1)*], C. Wenger[(2,3)], M. Ziegler[(1)]

[(1)] Energy Materials and Devices, Department of Materials Science, Kiel University, Kiel, Germany
[(2)] IHP – Leibniz Institute for High Performance Microelectronics, Frankfurt/Oder, Germany
[(3)] BTU Cottbus-Senftenberg, Cottbus, Germany
*Email: kni@tf.uni-kiel.de

The rapid advancement of artificial intelligence technologies is associated with an ongoing increase in energy consumption, driven by the growing computational demands of model training and inference. Energy efficiency can be significantly improved by transferring synaptic connections in neural networks to a hardware-based platform using memristive devices. At the same time, the switching principle of memristive devices is connected with stochastic processes involved in filament formation, which, in the context of neural networks, implies incompatibility with standard training methods based on gradient descent and the need to develop new algorithms.

To enable the construction of a scalable neural network with on-chip training based on variable weight switching, we propose a training algorithm that extend learning from mistakes, an approach to neural networks training whose origins lie in bio-inspired pruning [1]. For MNIST handwritten digits dataset classification [2], we use softmax regression as an example, that does not train sufficiently well with memristive weights using stochastic gradient descent alone, but in combination with learning from mistakes converge to high accuracy values. In this way, simulations of neural network training with weight variation during resistive switching, experimentally verified on a CMOS-integrated HfO_2-based resistive random-access memory (RRAM), establish a scalable framework for robust learning in memristive neural networks, bridging the gap between stochastic hardware and high-precision network performance.

[1] K. Nikiruy, et al., Scientific Reports, 18, 882-888, 2024.

[2] Deng L. IEEE Signal Processing Magazine. 29(6):141–2, 2012.