Overcoming phase-change material non-idealities by meta-learning for adaptation on the edge

Demirag, Yigit ; Dittmann, Regina ; Indiveri, Giacomo ; Neftci, Emre O

DOI: https://doi.org/10.29363/nanoge.neumatdecas.2023.043

Originally published at:
Demirag, Yigit; Dittmann, Regina; Indiveri, Giacomo; Neftci, Emre O (2023). Overcoming phase-change material non-idealities by meta-learning for adaptation on the edge. In: Neuromorphic Materials, Devices, Circuits and Systems, València, Spain, 23 January 2023 - 25 January 2023, Fundació Scito.
DOI: https://doi.org/10.29363/nanoge.neumatdecas.2023.043

Overcoming phase-change material non-idealities by meta-learning for adaptation on the edge

Yigit Demirag[1], Regina Dittmann[2], Giacomo Indiveri[1], Emre Neftci[2]

[1] The Institute of Neuroinformatics, UZH and ETH, Switzerland
[2] Peter Grünberg Institute, Forschungszentrum Juelich, Germany
yigit@ini.ethz.ch

Memristive crossbar arrays are built using scalable in-memory computing technologies. As such, these arrays can solve the von-Neumann bottleneck, which is the dominant factor of the energy consumption of neural network computations. They enable the deployment of large network models on compact embedded, mobile, edge devices for ultra-low power real-time inference applications [1].

However, in many scenarios, the post-deployment conditions might require a further adaptation in the network weights to compensate for domain data shifts due to sensory or environmental noise, gradual device degradation or simply fine-tuning to a new task that has not been seen during the offline training. The analog hardware non-idealities such as programming variability and non-linearity of conductance updates, the high power consumption of programming pulses, and the limited device endurance hinders the possibility of having power-efficient on-chip edge adaptation. This becomes further complicated with the data and memory-intensive requirements of gradient-based learning, such as training in batches and the lack of online and local credit assignment solutions.

In this work, we demonstrate that bi-level optimization offers an offline training solution to overcome the impact of such non-idealities for online learning on edge. We use Model-Agnostic Meta-Learning (MAML) [3] to pre-train a feedforward spiking neural network with accurate memristive synapse models to perform few-shot learning on standard regression benchmarks. Through the differentiable stochastic device model, the outer optimization loop learns a good initialization of the memristive conductances for the deployement that the negative impact of device programming during inner loop updates on the edge is suppressed. To achieve this, we use an empirical model of phase-change material (PCM) [2], which comprises all major forms of analog hardware non-idealities based on the measurements from 10,000 devices. The model includes the programming stochasticity and non-linearity as a function of device programming history, $1/f$ read noise, and temporal conductance drift. We improved the offline outer loop optimization performance by carefully selecting twice-differentiable surrogate spiking function [4] and additional hardware-aware training methods, such as noise-injection with straight-through estimators [5] and update thresholding [6].

Our results showed that a network trained with bi-level optimization can be finetuned to adapt new tasks on the edge by applying single blind programming pulses to only the output layer memristors of the neural network. We report competitive accuracies thanks to learning to mitigate write noise, read noise and drift of phase-change memory devices.

References

[1] Wan, W. et al. A compute-in-memory chip based on resistive random-access memory. Nature 608, 504–512 (2022).

[2] Nandakumar, S. R. et al. A phase-change memory model for neuromorphic computing. J Appl Phys 124, 152135 (2018).

[3] Finn, C., Abbeel, P. & Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Arxiv (2017).

[4] Stewart, K. M. & Neftci, E. O. Meta-learning spiking neural networks with surrogate gradient descent. Neuromorphic Comput Eng 2, 044002 (2022).

[5] Joshi, V. et al. Accurate deep neural network inference using computational phase-change memory. Nat. Commun.11, 2473 (2020)

[6] Demirag, Y., Frenkel, C., Payvand, M. & Indiveri, G. Online Training of Spiking Recurrent Neural Networks with Phase-Change Memory Synapses. Arxiv (2021).