Scaling Neural Networks On Accelerators With Procedural Connectivity Generation

Catherine M. Schöfmann[1,2], Jan Vogelsang[1,2], and Susanne Kunkel[1]

[1] Neuromorphic Software Ecosystems (PGI-15), Jülich Research Centre, Jülich, Germany

[2] RWTH Aachen University, Aachen, Germany

We introduce extensible simulation technology for spiking networks on massively distributed memory using Graphcore's Intelligence Processing Units (IPU) [1] and extend it for procedural generation of connectivity. This eliminates the need to store individual synapse data, enabling significantly larger model sizes and accelerating models that were previously not portable to memory-constrained hardware.

The computational nature of spiking simulations, frequent memory access with relatively sparse compute, causes them to be heavily bottlenecked by memory operations and communication when targeting conventional architectures. Distributed memory designs, popularized by the need for massively parallel and scalable processing for machine learning workloads, offer an alternative by tightly coupling processors with scratchpad memory (SRAM). Our simulator demonstrates the efficiency and scalability of an algorithm fully leveraging distributed memory.

Procedural connectivity generation heavily makes use of the IPU's hardware PRNGs on each core [2] in order to re-draw synapse targets, weights and delays from their respective distributions on the fly. This builds on top of a grouped and parallelizable delivery algorithm designed for spiking simulations on the IPU, which allows a controlled trade-off between performance and communication granularity. We implement a procedurally generated version of a benchmark model, the cortical microcircuit [3], and observe an over 75% reduction in memory footprint while matching the static model in dynamics and incurring minimal cost in speed.

[1] Z. Jia et al., arXiv, 2019.

[2] J. Hanlon and S. Felix, IEEE Trans. Comput., 72(5), 1518–1524, 2022.

[3] T. C. Potjans and M. Diesmann, Cereb. Cortex, 24(3), 785–806, 2012.