Towards energy-efficient hardware acceleration of memory-intensive event-driven kernels on a synchronous neuromorphic substrate

Abstract:

Spiking neural networks are increasingly becoming popular as low-power alternatives to deep learning architectures. To make edge processing possible in resource-constrained embedded devices, there is a requirement of reconfigurable neuromorphic accelerators that can cater to various topologies and neural dynamics typical to these networks. Subsequently, they also must consolidate energy consumption in emulating these dynamics. Since spike processing is essentially memory-intensive in nature, majority of the system's power consumption can be reduced by eliminating redundant memory traffic to off-chip storage that holds the large synaptic data of the network. In this work, we first present a digital synchronous neuromorphic accelerator that can emulate different types of spiking neurons and network topologies for efficient inference. It is functionally verified on a set of benchmarks that vary significantly in topology and activity while solving the same underlying task. By studying the memory access patterns, locality of data and spiking activity, we establish the core factors that limit conventional cache replacement policies from performing well. We, then propose a domain-specific memory management scheme that exploits our use-case to attain visibility of future data-accesses in the event-driven simulation framework. To make it even more robust to variations in network topology and activity of the benchmark, we further propose static and dynamic network-specific enhancements to adaptively equip the scheme with more insight. The strategy is explored and evaluated with the set of benchmarks using a software simulation of the accelerator and an in-house cache simulator. In comparison to conventional policies, up to 23% more reduction in net power consumption is observed.