Spike-based Approaches to Efficient Sparse Modeling and Visual Processing

Name Takumi Kuwahara
Laboratory Computing Architecture
Supervisor Prof. Yasuhiko Nakashima

Abstract

Spiking sparse modeling is a key technology for event-driven, energy-efficient neuromorphic computing. The Spiking Locally Competitive Algorithm (S-LCA) is a low-power, low-latency LASSO solver, yet its long convergence timestep prevents integration with modern few-timestep spiking neural networks. This dissertation addresses this limitation by advancing spiking sparse modeling toward few-timestep operation. First, we propose Modern Spiking LCA (MS-LCA), a directly trained variant of S-LCA using Backpropagation Through Time. MS-LCA reduces timestep by about 250× while maintaining accuracy comparable to analog LCA, enabling low-latency reconstruction of static images. Next, we introduce Dynamic MS-LCA (DMS-LCA), which extends MS-LCA with temporal dictionaries to process event-based data such as N-MNIST and POKER-DVS. DMS-LCA offers controllable sparsity and lowers computational cost. On POKER-DVS, reconstruction accuracy slightly deteriorates compared with S-LCA, yet synaptic operations (SynOps) are reduced by 92%. We further design a two-stage hierarchical model using MS-LCA and DMS-LCA with residual connections and staged loss functions. This model reduced CIFAR-10 reconstruction error by 42% and SynOps by 28% relative to S-LCA.

In the second part of this dissertation, we investigate a hardware-oriented approach using a fusion synapse that combines a memristor and a capacitor. This structure allows both a wide weight range and low power consumption by representing the synaptic weight with the time constant of the circuit. Network parameters are converted into conductances within circuit simulations. The proposed system achieves 95% accuracy on MNIST at 640 nJ per inference, suggesting potential benefits compared with digital implementations when considering Dennard scaling.

Overall, this dissertation presents advances in both software algorithms and hardware architectures for power-efficient neuromorphic computing.