ScRRAMBLe: Block-Sparse Deep Learning Architecture for Analog Compute-in Memory Accelerators

Vikrant Jaltare[1,2], Rommani Mondal[1,2], Leif Gibb[2], Jiahao Song[2], Johannes Leugering[3] and Gert Cauwenberghs[1,2]
[1] Shu Chien-Gene Lay Department of Bioengineering, UC San Diego, [2] Institute for Neural Computation, UC San Diego, [3] Neuromorphic Compute Nodes (PGI-14)
Forschungszentrum Jülich, Germany

for more info!

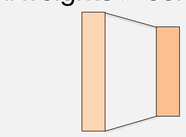
Promise of Analog Compute-in Memory (CIM)*

- Matrix-vector multiplication in $\mathcal{O}(1)$
- High-density of conductances ⇒ large capacity of parameters.
- Non-volatile storage.
- Analog multi-level weights ⇒ inexpensive high-precision weight quantization.

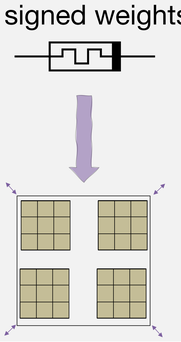
** applies to conductance-based/memristive technologies*

Computational Bottlenecks

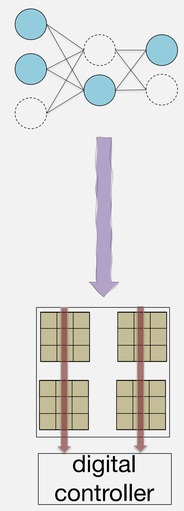
#weights > core-size



positive conductance // signed weights



unstructured sparsity

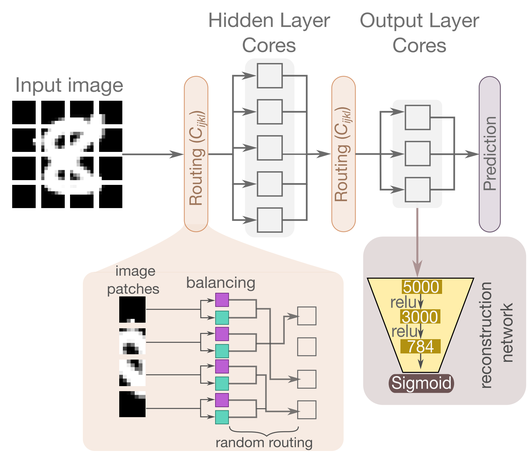


area on Si cost **communication cost**

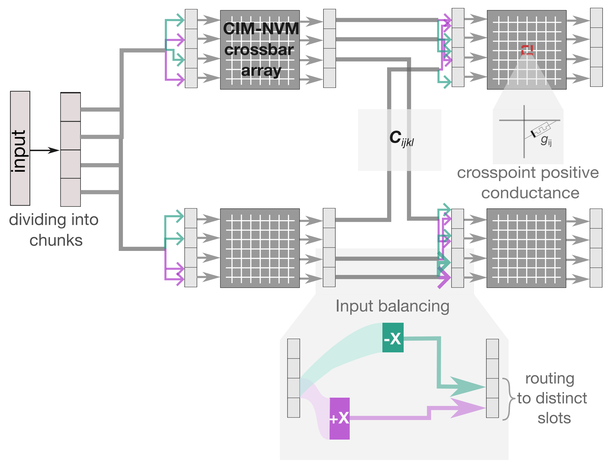
ScRRAMBLe leverages block-sparsity to fully utilize dense cores and reduce inter-core communication.

Neural Network and Hardware Co-Design

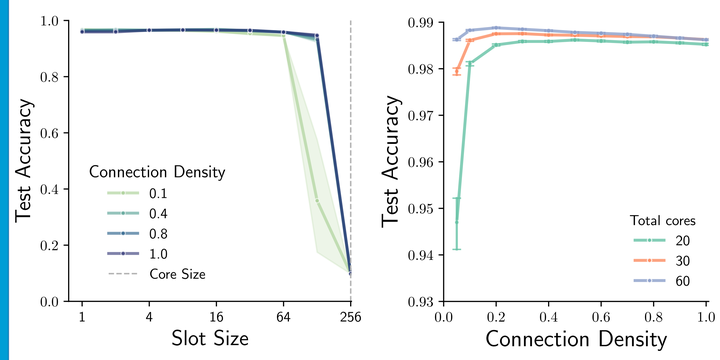
Network Architecture



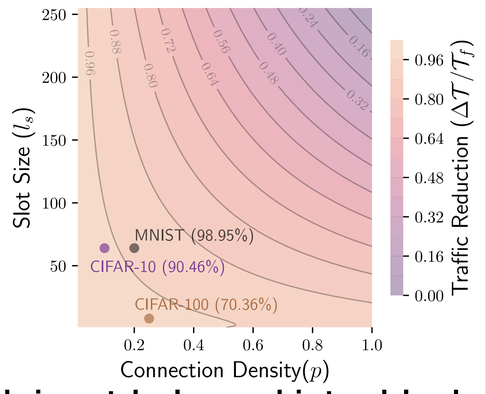
Hardware Mapping



Block-Sparsity and Performance

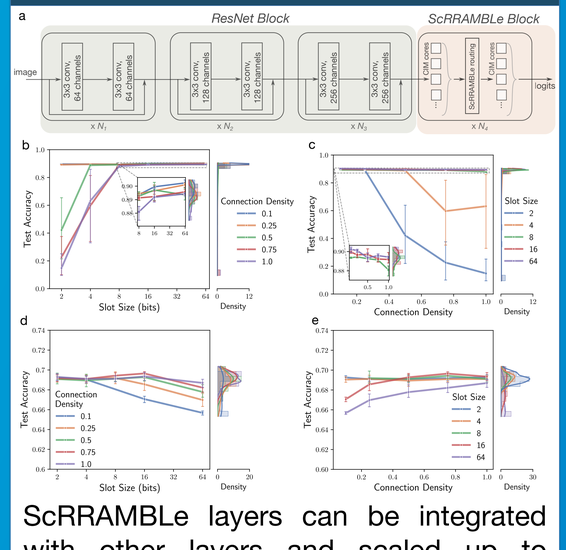


Intercore communication



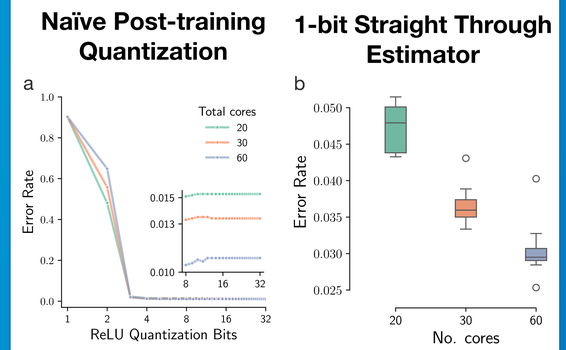
Block-sparse neural networks together with input-balanced inter-block routing can replace computationally expensive fully-connected layers in analog CIM accelerators while implementing a signed weight per positive crossbar conductance.

Modular Scaling



ScRRAMBLe layers can be integrated with other layers and scaled up to deeper neural networks.

Quantization Robust Training



Quantization Aware Training methods like Straight Through Estimator rescue performance in networks with extreme activation quantization. Future studies are needed to test scalability of this approach.

In press
npj | unconventional computing

Funding Sources

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Shu Chien-Gene Lay Department of Bioengineering

INTEGRATED SYSTEMS NEUROENGINEERING