Neuromorphic Computing

Wesley Tian

Overview

- Neuromorphic Computing
- Memristors 101
- Applications
- Main Literature
  - TraNNsformer
- Additional Literature
  - Neurogrid
  - Neuromorphic silicon neurons and large-scale neural networks

Memristors 101

Applications

- Non-volatile random access memory (NVRAM)
- Transistor replacements
- Revisit analog computation
- Learning circuits
- AI
- Automotive/Industrial/Medical
- Consumer Electronics
- Data Center
- IoT
- Mobile Computing

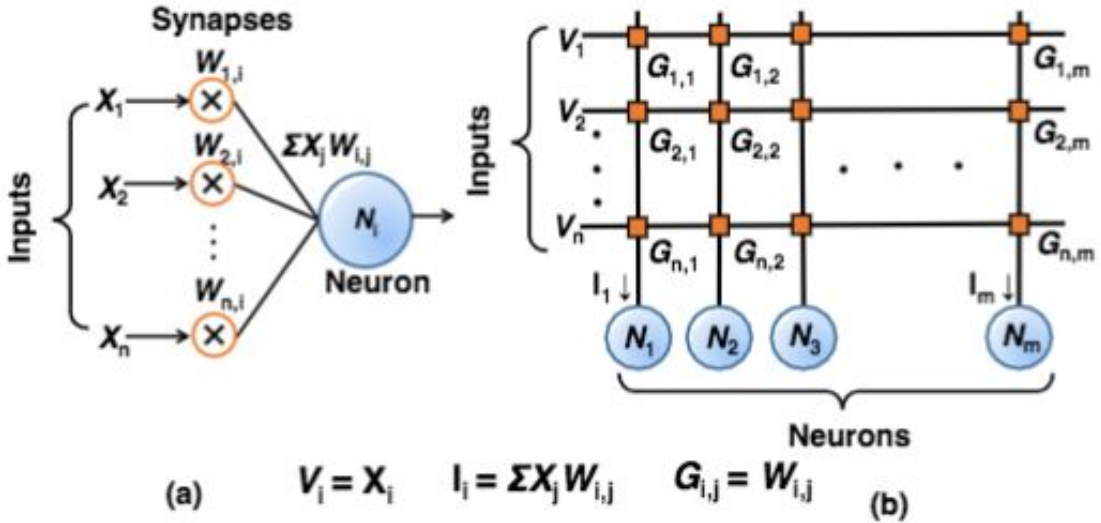
TraNNsformer: Neural Network Transformation for Memristive Crossbar based Neuromorphic System Design

Aayush Ankit, Abhronil Sengupta, Kaushik Roy
School of Electrical and Computer Engineering, Purdue University
26 Aug 2017
Accepted in IEEE/ACM ICCAD 2017

Contents

- **Introduction** and related work
- **Present** Size-Constrained Iterative Clustering (SCIC) algorithm
- **Proposal** of TraNNsformer - an Integrated Training Framework
- **Evaluation** of proposed methodology
- **Analysis** of resulting energy and energy benefits
- **Conclusion**

I. Introduction and Related Work



$V_i = X_i$ $I_i = \Sigma X_j W_{i,j}$ $G_{i,j} = W_{i,j}$

Fig. 1: (a) A two-layered MLP based Neural Network (b) Neural Network mapped to Memristive Crossbar Array (MCA)

I. Introduction and Related Work

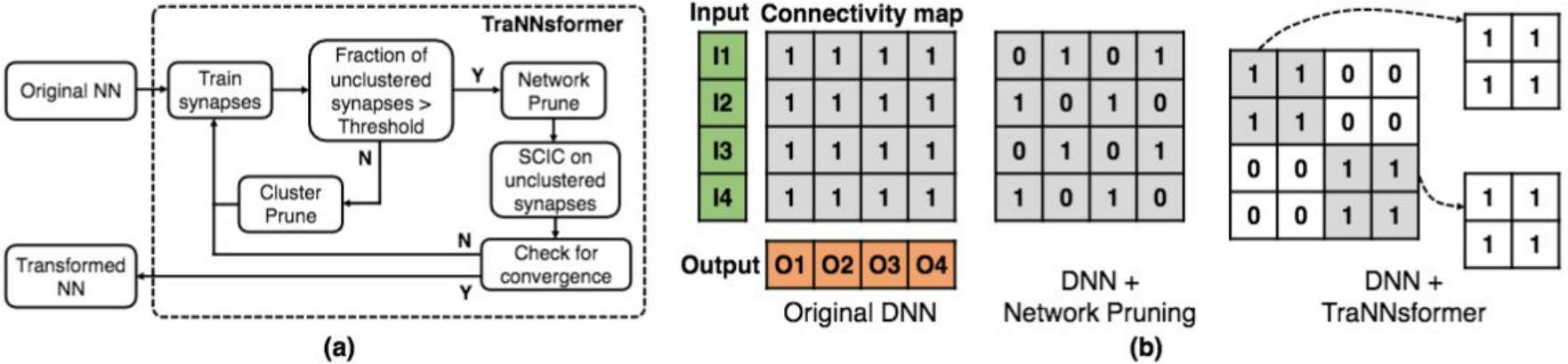


Fig. 2: (a) Logical Flow Diagram of TraNNsformer Framework. The original DNN architecture during training undergoes clustering to form regions that can be mapped onto MCAs with high utilization factors, while pruning the connections that don't contribute to cluster formation. (b) Toy example to illustrate the impact of Network pruning and TraNNsformer on a DNN connectivity matrix. Network pruning leads to irregular sparsity that cannot be mapped directly onto MCAs. TraNNsformer forms smaller clusters that can be mapped onto MCAs. Note that 1/0 only represents a connection being present and not the actual value of the weight.

III. TraNNsformer Framework

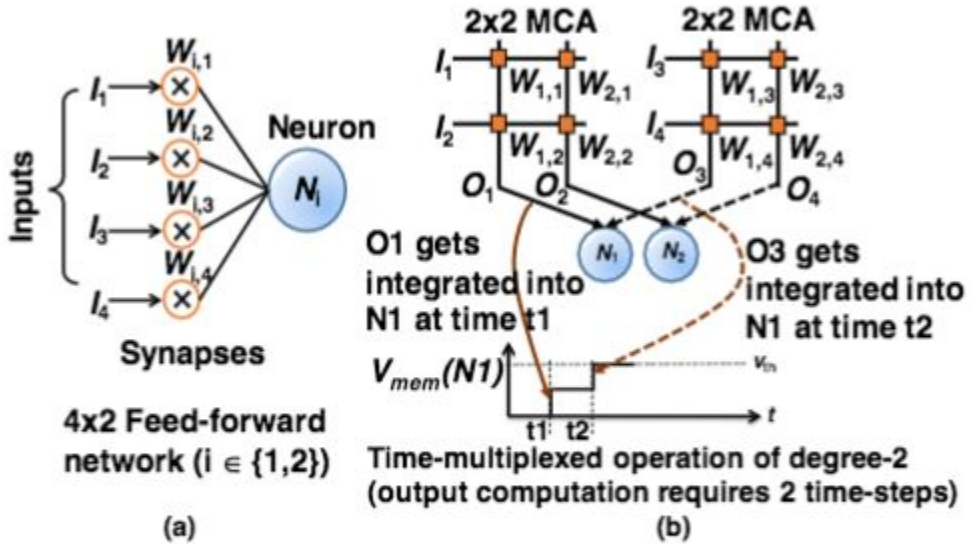


Fig. 3: (a) A feed-forward neural network with neuron fan-in of 4 (b) Mapping the 4 fan-in neurons using a 2×2 MCAs

IV. Experimental Methodology

Application	*Dataset*	*Layers*	*Neurons*	*Synapses*
Digit Recognition	MNIST	4	3194	2392800
House Number Recognition	SVHN	5	4634	4120800
Object Classification	CIFAR-10	6	5834	5560800

Fig. 4: MLP based SNN benchmarks

IV. Experimental Methodology

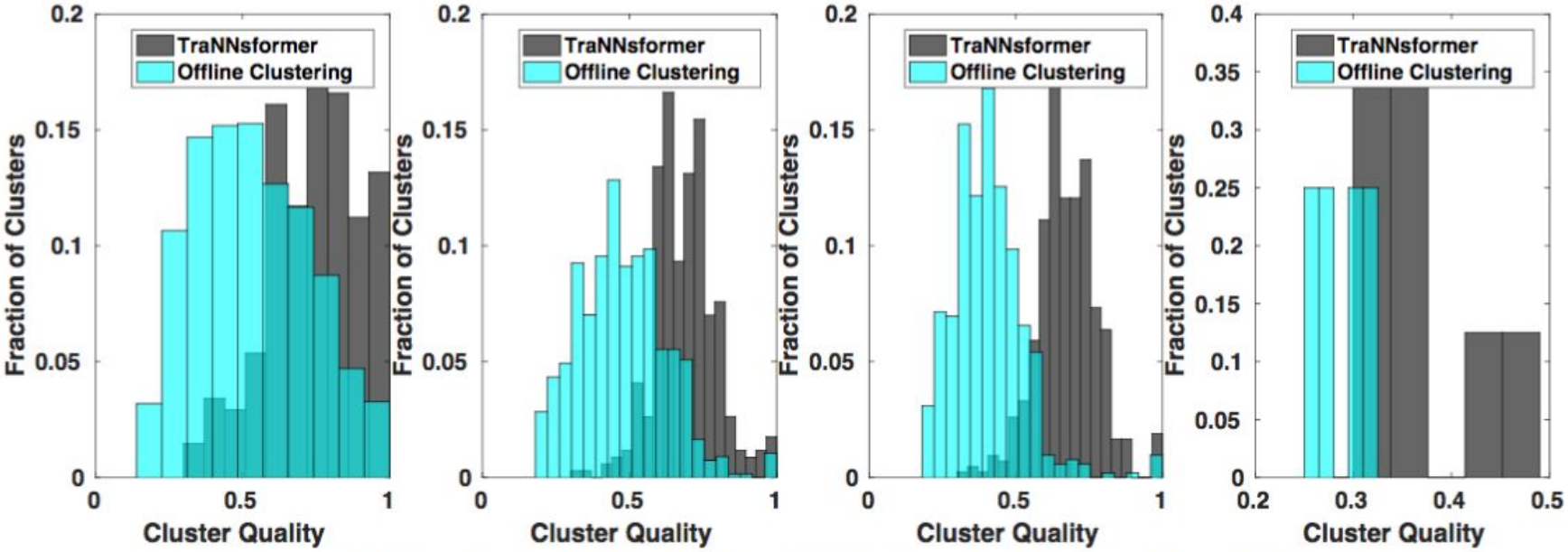


Fig. 5: Comparison of crossbar utilization (cluster quality) between Offline Clustering and TraNNsformer

Algorithm level analysis

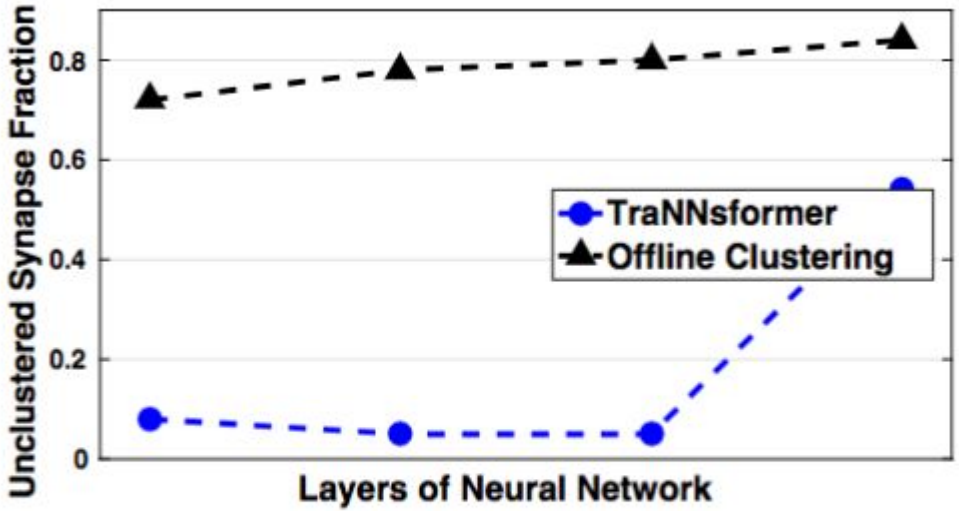


Fig. 6: Comparison of fraction of unclustered synapses between Offline Clustering and TraNNsformer (the data points correspond to the layers of DNN). Note that the last fully connected layer consists of a small fraction of synapses (<1%), thereby having insignificant effect on overall unclustered synapse comparison.

Algorithm level analysis

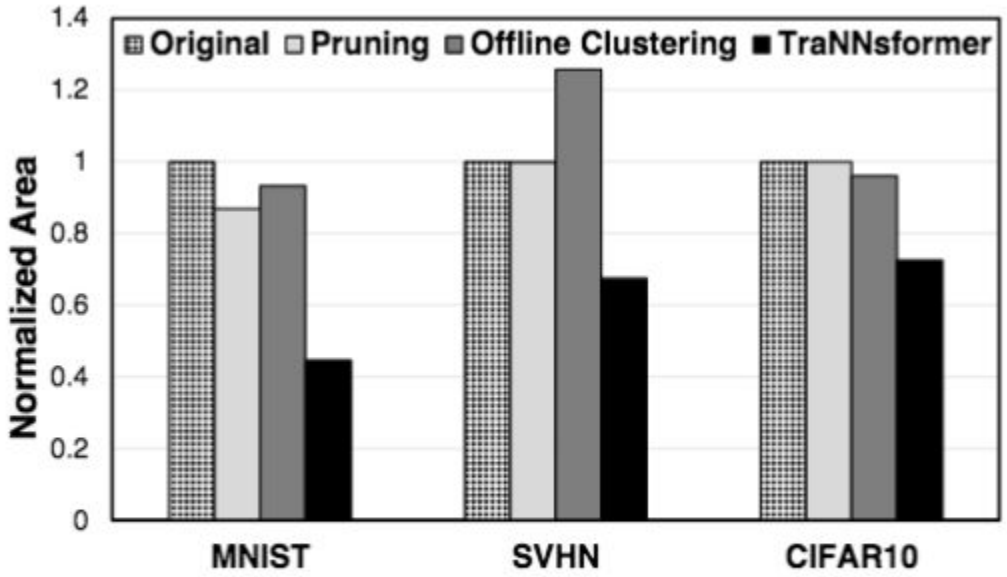


Fig. 7: Comparison of area consumption on MCA based architecture for different DNN training approaches

Comparison of energy consumption on MCA based architecture for different DNN training approaches

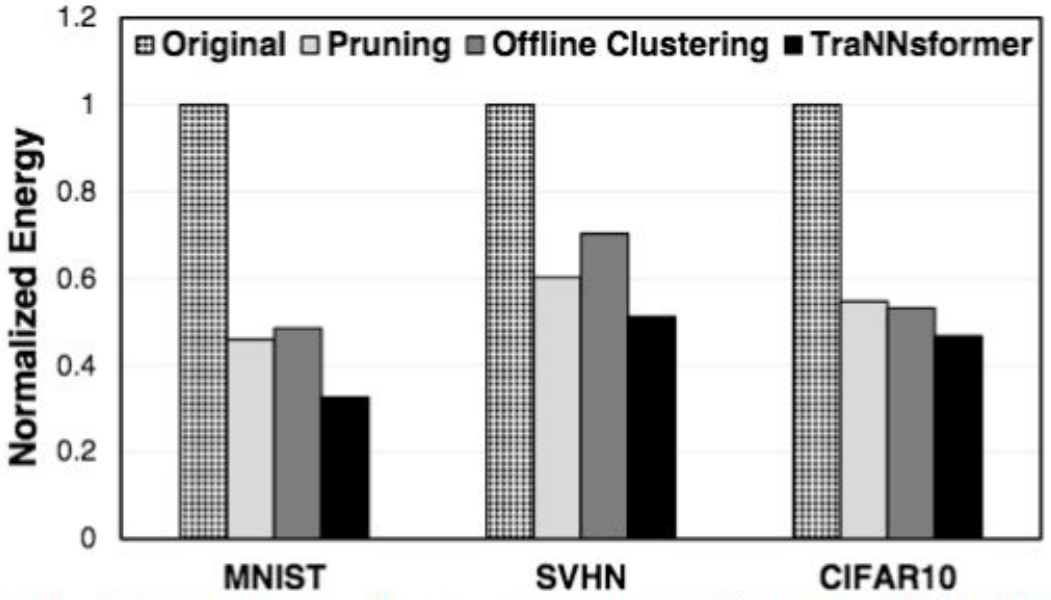


Fig. 8: Comparison of energy consumption on MCA based architecture for different DNN training approaches

Energy comparison of CMOS based general-purpose architectures

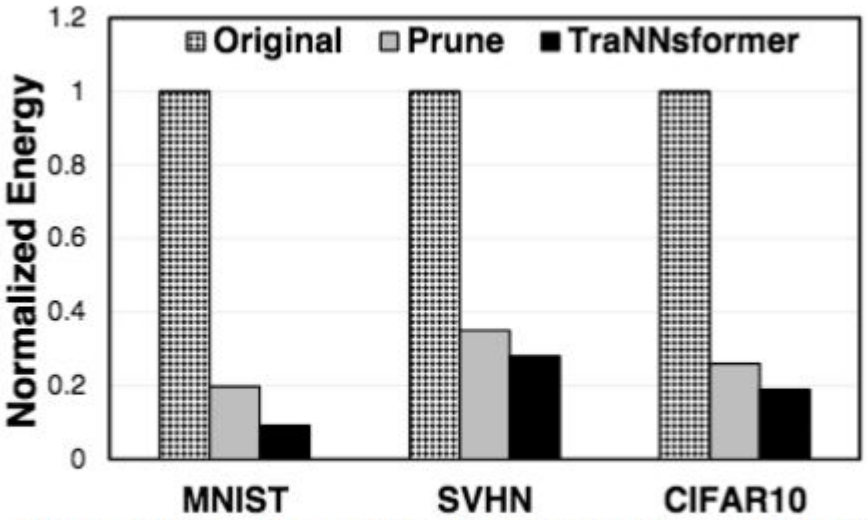


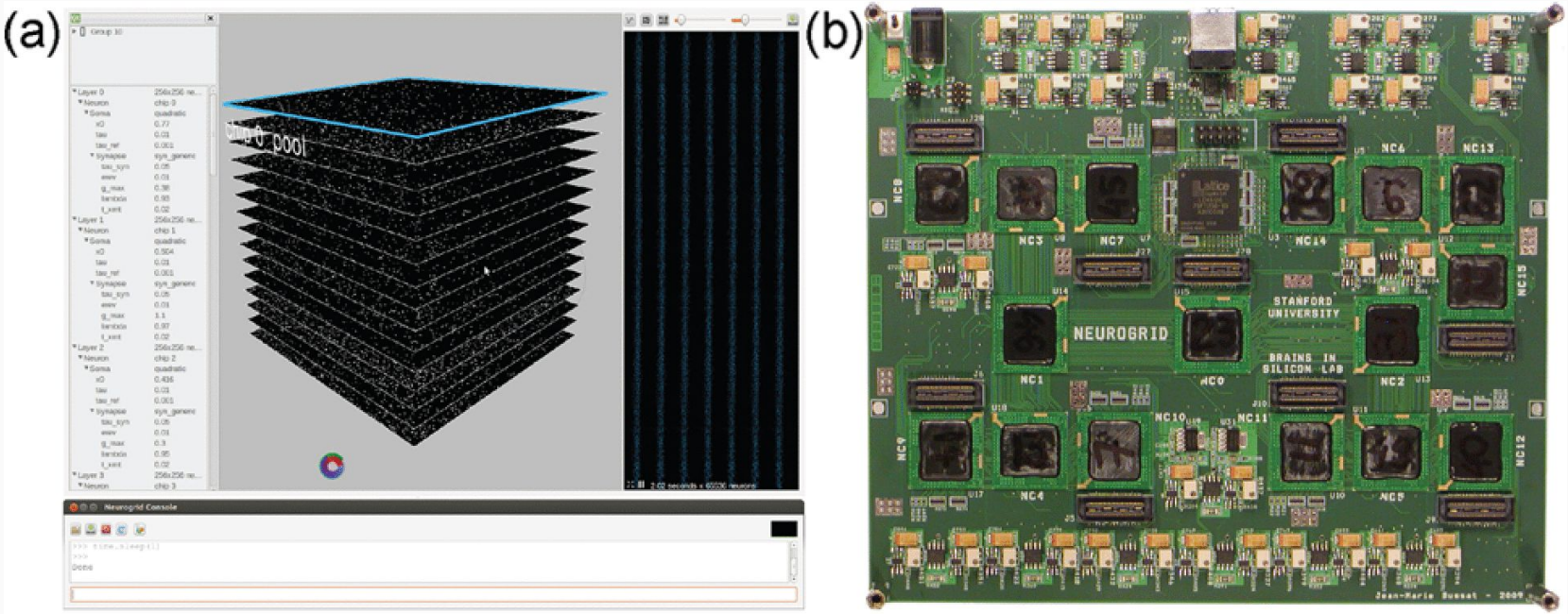
Fig. 9: Comparison of energy consumption on CMOS based general-purpose architecture for different DNN training approaches

Additional Literature

Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations

Benjamin
Dept. of Electr. Eng., Stanford Univ., Dept. of Bioeng., Stanford Univ., In-Depth, Inc., Intel Corp., Mad Street Den, Apple Comut., Almaden Res. Center, IBM, 24 April 2014
Proceedings of the IEEE (Volume: 102, Issue: 5, May 2014)

Neurogrid



(a) GUI: Enables a user to change his or her model parameters (left), view spike activity in the model's various layers (middle), plot spike rasters from a selected neural layer (right), and enter commands (bottom). (b) Board: Each neural layer is simulated by up to 256 x 256 silicon neurons on each of 16 Neurocores integrated on a 6.6 x 7.5 in^2 board.

Simulating Large-Scale Neural Models

- A personal computer simulates a mouse-scale cortex model (2.5e6 neurons) 9000 times slower than a real mouse operates
- *Uses 40,000 more power (400W vs 10mW)*
- Simulating human-scale cortex model (2e10 neurons), the Human Brain Project's goal is projected to require an exascale supercomputer (10^{18} flops)
- *As much power as a quarter-million households (0.5 GW)*
- Note: China's Sunway TaihuLight has a benchmark rating of 93 PFLOPS (9.3e16)

Summary

- Made it possible to simulate a million neurons with billions of synaptic connections in real time-for the first time-using 16 Neurocores integrated on a board that consumes 3W
  - Emulated all neural elements except the soma with shared electronic units → maximized the number of synaptic connections
  - Realized all electronic circuits except those for axonal arbors in an analog manner → maximized energy efficiency
  - Interconnected neural arrays in a tree network → maximized throughput

Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities

Chi-Sang Poon[1] and Kuan Zhou[2]
[1]Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Intel Corporation, Hillsboro, OR, USA

Neuromorphic Silicon Neurons vs Digital Neural Simulations

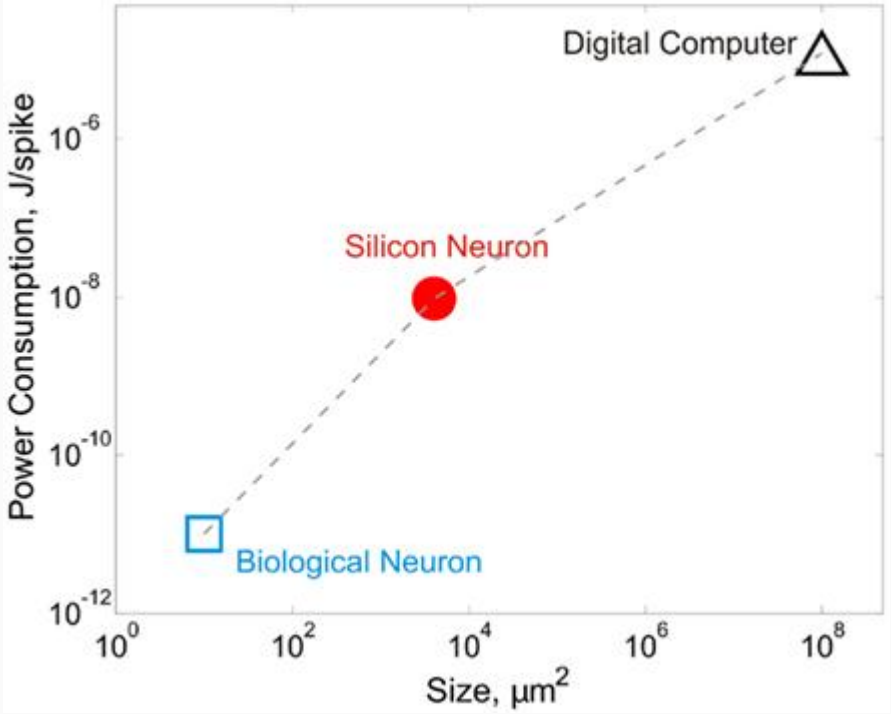


Figure 1. Biological and silicon neurons have much better power and space efficiencies than digital computers.