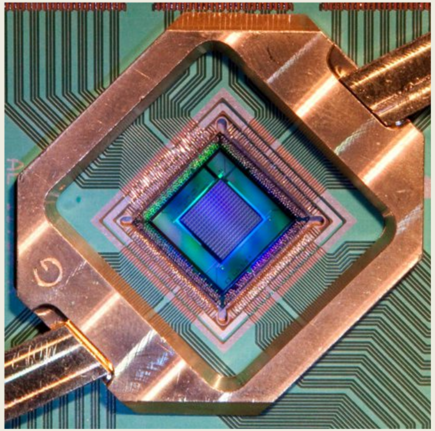
Biologically Inspired Sparse Coding on a Neuromorphic Device

Kyle Henke // LANL CCS-3 and the University of New Mexico
Benjamin Migliori // LANL CCS-7
Garrett Kenyon // LANL CCS-3

Special thanks to collaborators:
Nga Nguyen // LANL CCS-3
Carleton Coffrin // LANL A-1

The original problem: posing the same challenge to different computing substrates

D-Wave quantum annealer



Solves a spin glass problem by finding lowest energy state.

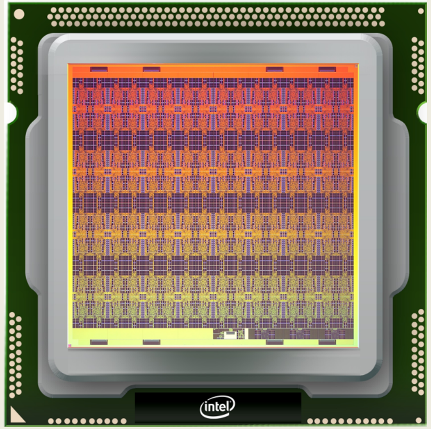
Initializes system into a superposition of all spins, and then anneals on problem Hamiltonian.

In theory, system tunnels down to global minimum.

Solution is lowest *observed* energy state.

Operates in vacuum 180x colder than deep space; limited to ~64 logical cubits.

Intel Loihi neuromorphic processor



Simulates biological spiking neurons as compartmental models.

Fully asynchronous integrate-and-fire with synaptic plasticity.

Solution reads out as neural spikes.

Requires casting constraint satisfaction problems as anatomy problems; leans heavily on bioinspiration.

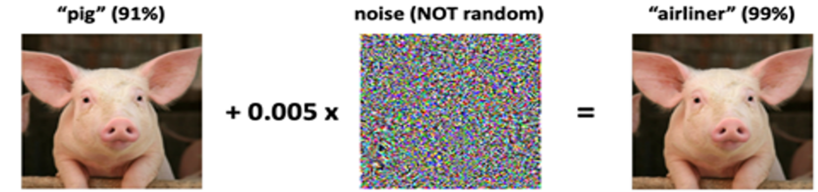
Novel Idea: Demonstrate the implementation of *the same optimization problem* with *the same dataset* on these fundamentally different architectures.

Why Sparse Coding?

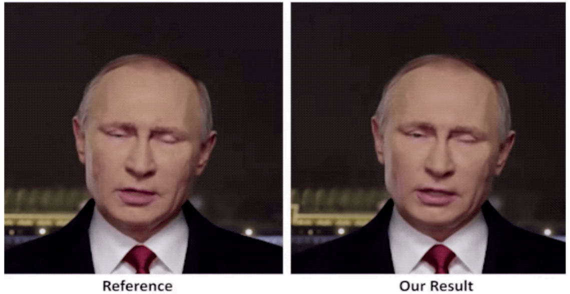
Weaknesses of Current Neural Network Architectures

1) Poor ability to generalize or transfer models to new data.
2) Vulnerability to adversarial attacks.
3) Emergence of deep fake media.
4) End of Moore’s Law reduces ability to scale in future.

Solutions?

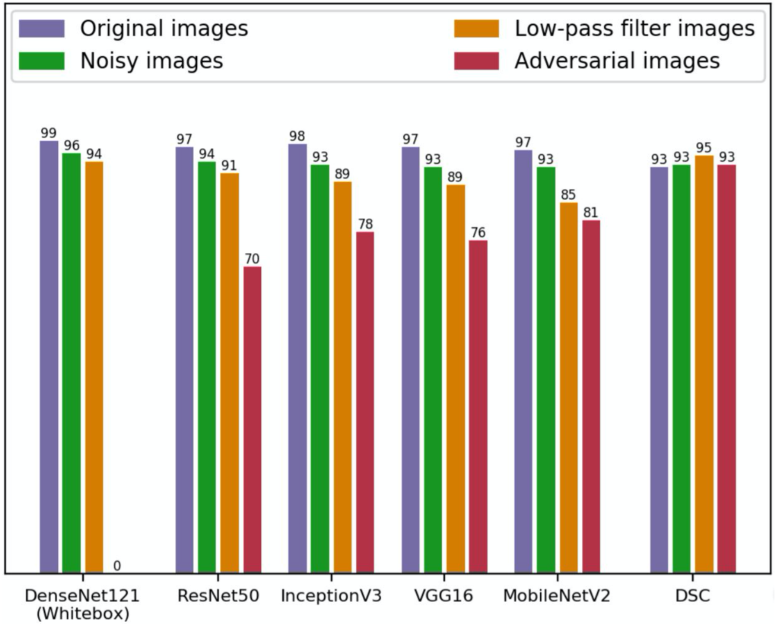


(Szegedy Zaremba Sutskever Bruna Erhan Goodfellow Fergus 2013)



Generative Adversarial Network (GAN) Deep Fake Video: Chebbi 2018

Deep Sparse Coding (DSC) Results vs Adversarial Examples



State of the art deep convolutional networks performance with different alterations to inputs. Adversarial examples trained on DenseNet121. Springer et. al 2018

Take Away : Sparse coding has been shown to be robust against common types of adversarial attacks and can also separate real from fake videos better than many existing techniques.

The specific problem: Solving Convex and Non-convex optimization on spiking device

Unsupervised learning w/binary sparse coding and optimization with Locally Competitive Algorithm (LCA)

An efficient, bio-inspired way of representing input in higher dimensions.
Given an overcomplete basis, sparse coding algorithms seek to identify the minimal set of dictionary generators (ϕ) that most accurately reconstruct each input.
NP-hard for L0 norm – non-convex
Effective L-P norm with Spiking LCA on Loihi - convex

$$E(\vec{\boldsymbol{I}},\vec{\boldsymbol{a}}) = \min_{\{\vec{\boldsymbol{a}},\boldsymbol{\phi}\}} [\frac{1}{2}||\beta\vec{\boldsymbol{I}}-\boldsymbol{\phi}\vec{\boldsymbol{a}}||^2+\lambda||\vec{\boldsymbol{a}}||_{0,p}]$$

Sparse coding as an Ising model/QUBO

Loihi can solve the spin glass problem described by the Hamiltonian:

$$H(\vec{\boldsymbol{h}},\boldsymbol{Q},\vec{\boldsymbol{a}}) = \sum_{i}^{N_q} h_i a_i + \sum_{i<j}^{N_q} Q_{ij} a_i a_j$$

Where $\vec{a}$ are binary coefficients.

By defining:

$$\vec{\boldsymbol{h}} = -\phi^T\vec{\boldsymbol{I}} + (\lambda + \frac{1}{2})$$

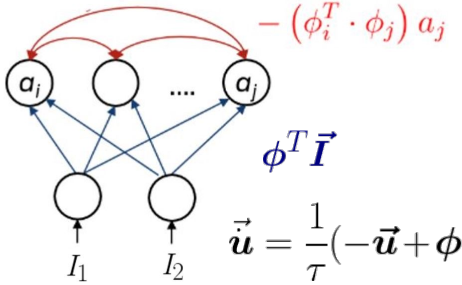
$$\boldsymbol{Q} = \phi^T\phi$$

we transform *H* into a sparse coding problem, where $\vec{a}$ represents the minimum energy sparse solutions.

Nguyen et al, Arxiv 2016

Sparse coding as lateral inhibition and competition with spikes

Bio-inspired neurons **send binary signals as spikes** in response to current - and can excite or inhibit each other.
Cast inputs as ***weighted injected current*** **(**$\phi^T\vec{\boldsymbol{I}}$**)**, and outputs as ***spike rates*** **(**$\vec{\boldsymbol{a}}$**)**


$-\left(\phi_i^T \cdot \phi_j\right) a_j$
a_i
....
a_j
$\phi^T\vec{\boldsymbol{I}}$
I_1
I_2

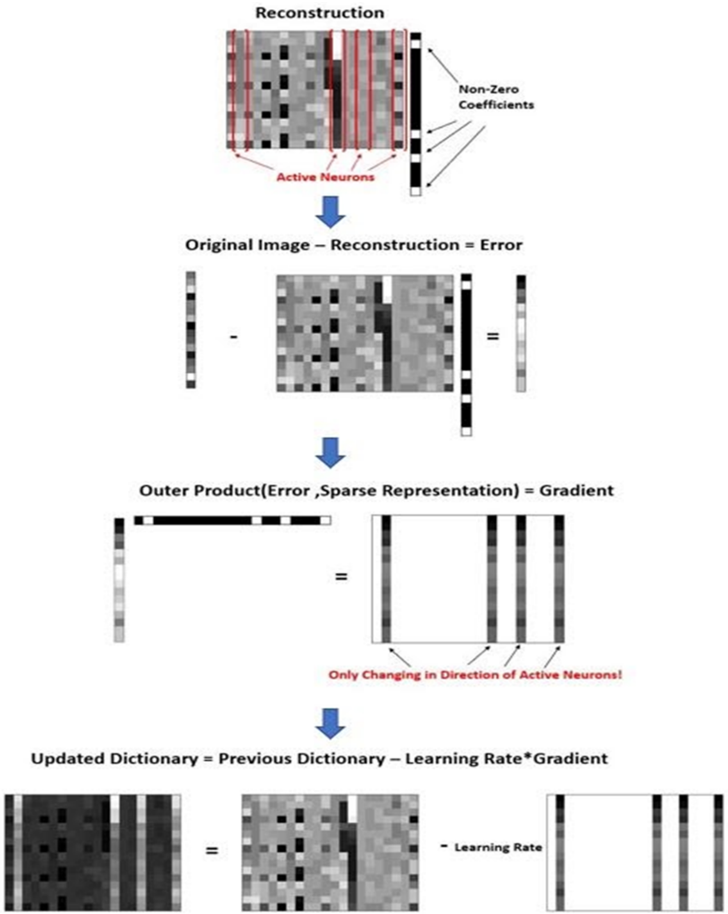
$$\dot{\vec{\boldsymbol{u}}} = \frac{1}{\tau}(-\vec{\boldsymbol{u}}+\phi^T\vec{\boldsymbol{I}}-\phi^T\phi\cdot\vec{\boldsymbol{a}}+T_\lambda(\vec{\boldsymbol{u}}))$$

Assume existence of input/output transfer function $\vec{\boldsymbol{a}} = T_\lambda(\vec{\boldsymbol{u}})$
Each neuron represents a generator; when active, inhibit other generators.
Neurons *compete and converge* to a sparse code over time.

Davies et al, 2017

Result 1: Bio-inspired LCA sparse coding allows interesting problems to be implemented on neuromorphic and quantum computers.

Before we get final solutions ... Dictionary optimization w/stochastic gradient descent (SGD)and Hebbian learning

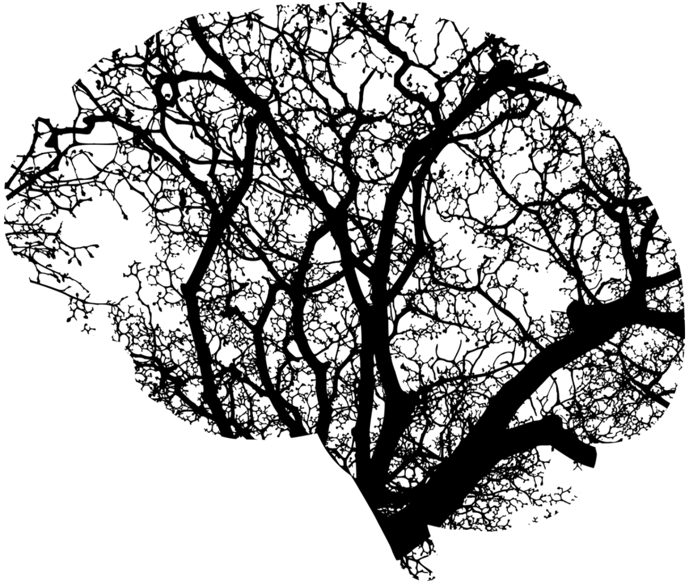


Original Image – Reconstruction = Error

Outer Product(Error ,Sparse Representation) = Gradient

Updated Dictionary = Previous Dictionary – Learning Rate*Gradient

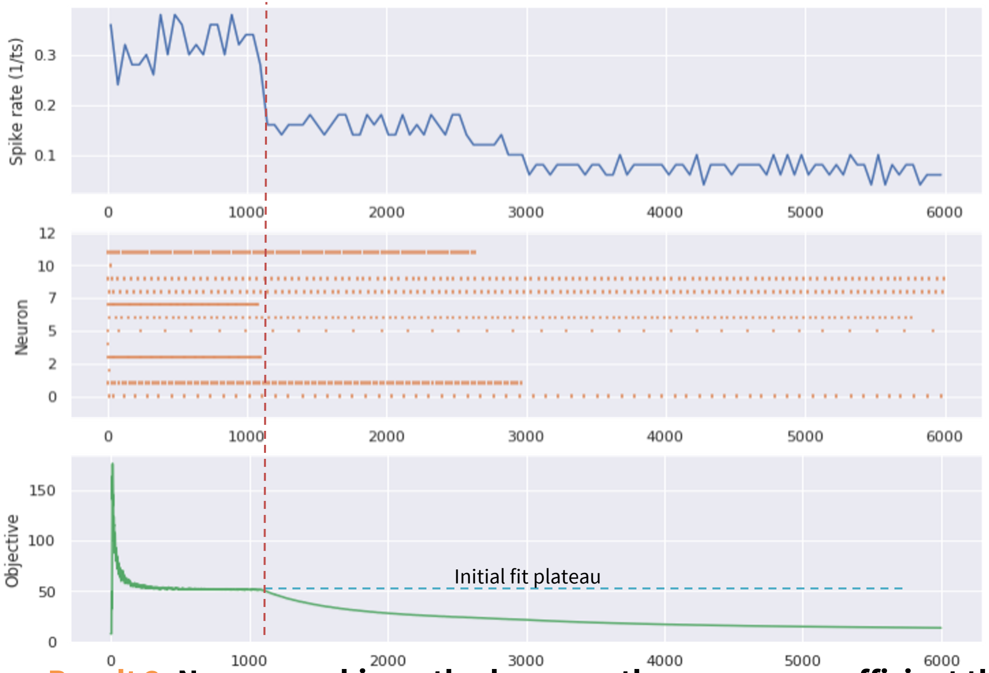
$$E(\vec{I},\vec{a}) = \min_{\{\vec{a},\phi\}} [\frac{1}{2}||\beta\vec{I}-\phi\vec{a}||^2+\lambda||\vec{a}||_{0,p}]$$

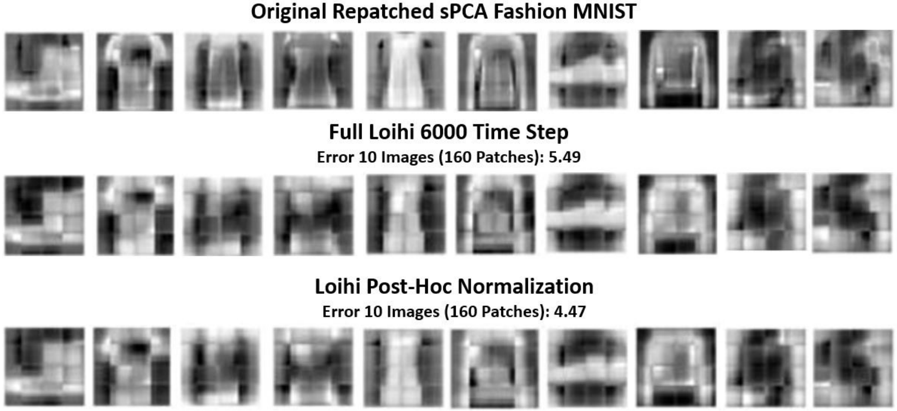


Biologically accurate local Hebbian rule allows for the unsupervised learning of an optimal set of overcomplete basis vectors to best represent the data set.

Neuromorphic local computing performance using spiking LCA

- Intel Loihi sparse coding with 6000 timesteps in simulation = 128ms simulation clock = ~13 seconds wall clock .
- Runs on 1.2 Watts!
- Most of the learning takes place in the first 100-200 simulation timesteps, but many active features.
- Full Loihi solutions are effectively an L-p norm with SPIKE RATES as the solution vector.





Original Repatched sPCA Fashion MNIST

Full Loihi 6000 Time Step

Error 10 Images (160 Patches): 5.49

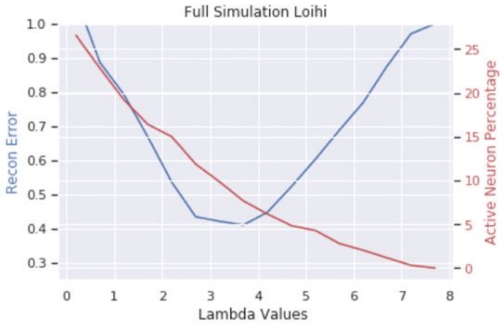
Loihi Post-Hoc Normalization

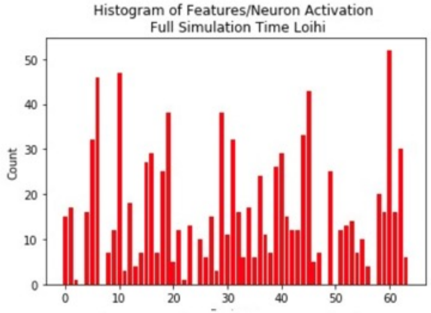
Error 10 Images (160 Patches): 4.47

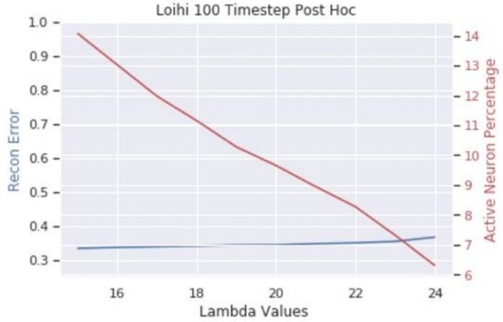
Result 3: Neuromorphic methods are vastly more power efficient than classical methods and require early in time post—hoc normalization for a more uniform solution vector to better approximate desired binary L-0 norm solutions found on the D-Wave.

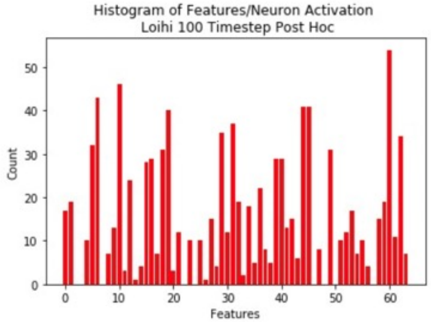
Post hoc tuning

- Post hoc normalization keeps many more sparse coefficients for solutions because active neurons haven't been dampened out over the competition.
- Amplify lambda sparsity parameter (λ) to achieve save sparsity level as full simulation.





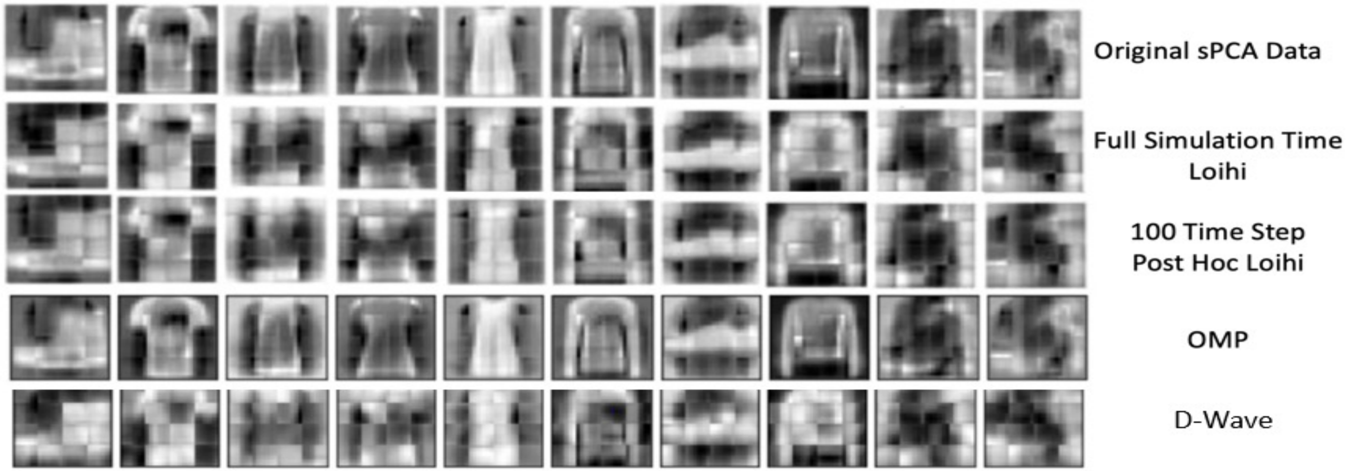


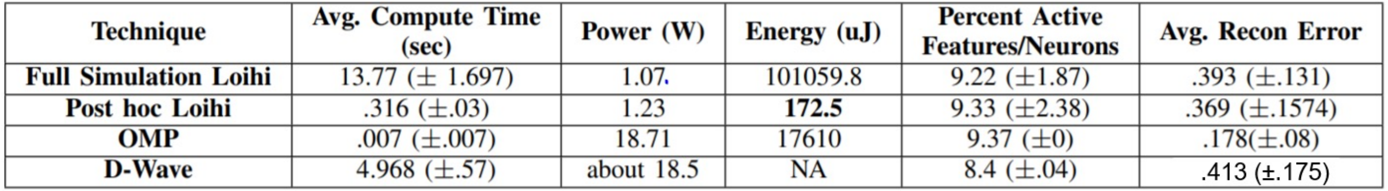


Result 7: After increased sparsity parameter, post hoc Loihi has similar feature activation to the full simulation counterpart.

Published Results

Remember: 1) D-Wave WAS only true binary sparse code solver.
2) Orthogonal Matching Pursuit (OMP) is greedy w/continuous solutions

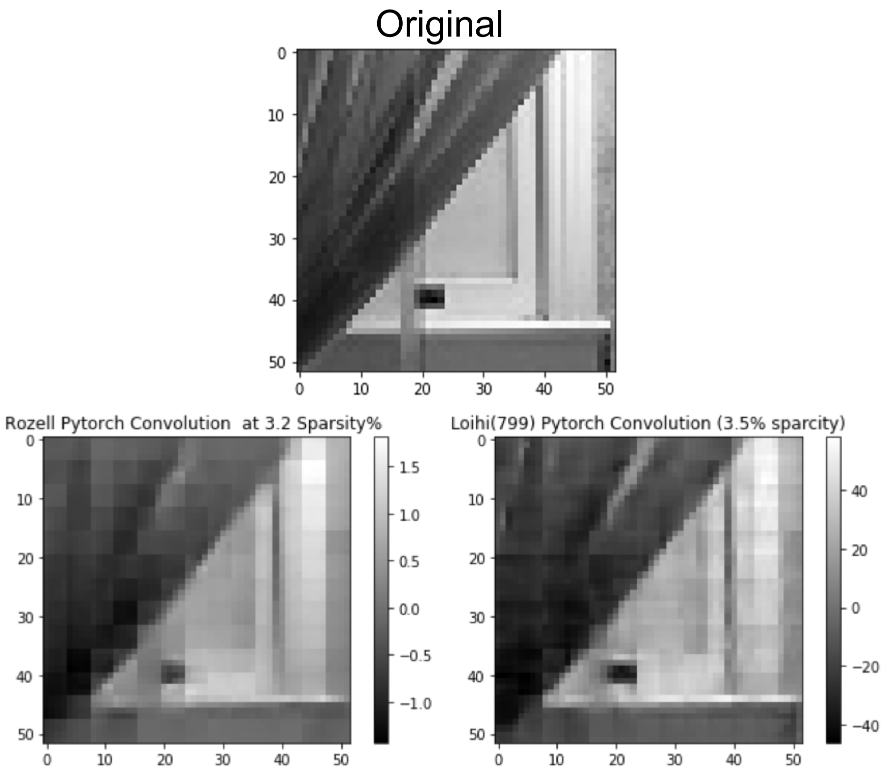


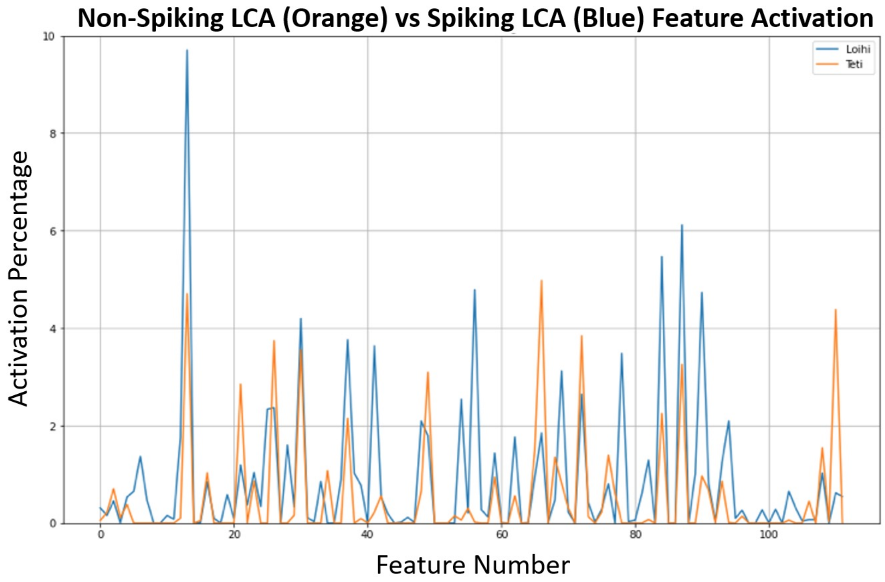


Technique	Avg. Compute Time (sec)	Power (W)	Energy (uJ)	Percent Active Features/Neurons	Avg. Recon Error
Full Simulation Loihi	13.77 (± 1.697)	1.07.	101059.8	9.22 (±1.87)	.393 (±.131)
Post hoc Loihi	.316 (±.03)	1.23	**172.5**	9.33 (±2.38)	.369 (±.1574)
OMP	.007 (±.007)	18.71	17610	9.37 (±0)	.178(±.08)
D-Wave	4.968 (±.57)	about 18.5	NA	8.4 (±.04)	.413 (±.175)

Result 8: Post hoc Loihi gives 50x speed up and often better solutions than the full simulation Loihi counterpart at a fraction of the energy consumption and power of quantum and classical algorithms.

First demonstration of non-spiking and spiking LCA equivalence for convolutional sparse code



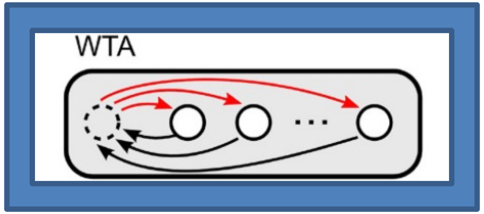


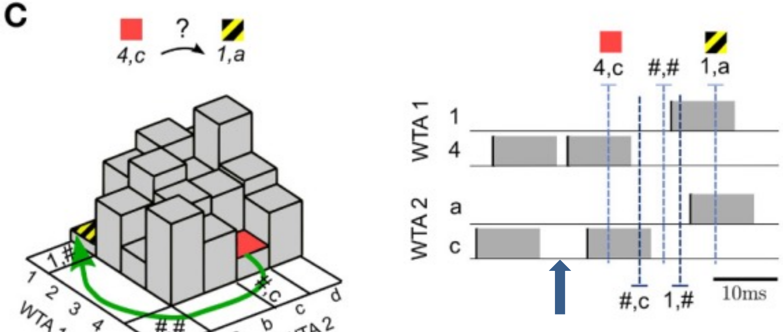
Result 2: **Lower energy spiking Loihi can solve equivalent classical bio-inspired LCA**

Constraint Satisfaction Problems (CSP) on Loihi

1. **Previous work by Hopfield (1986) and others showed CSPs can be solved with non-spiking neural networks**
2. **Networks of spiking stochastic neurons with SYMMETRIC weights are related to Boltzmann distributions.**
3. **By representing ‘Principle Neurons’ as asymmetric weighted networks of spiking neurons, we can better approximated the lowest energy orientation of a CSP when the network reaches steady state.**
4. **Spike based approaches allow for networks to navigate around high energy barriers because of refractory period.**

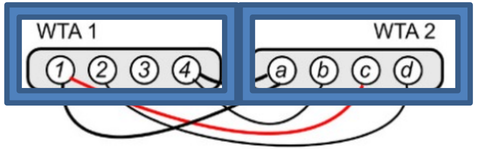
Principle Neuron represented by a ‘Winner Take All’ (WTA) Network





Refractory Period

WTA/Principle Neurons connected by symmetric weights



Searching All Possible Configurations to find the lowest energy

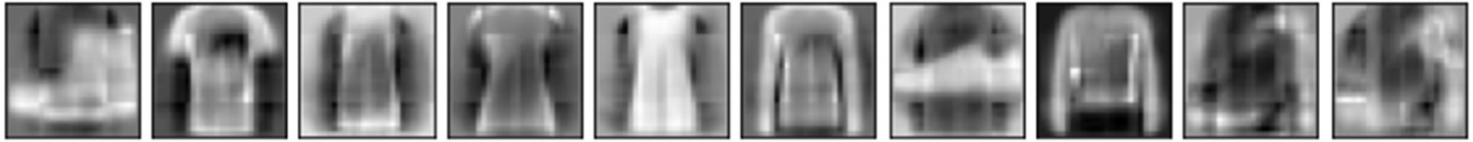
Boltzmann distribution underline all QUBO problems, so they fall under CSP category and can also be solved on LOIHI

QUBO Sparse Coding of Dimensionally Reduced Fashion MNIST

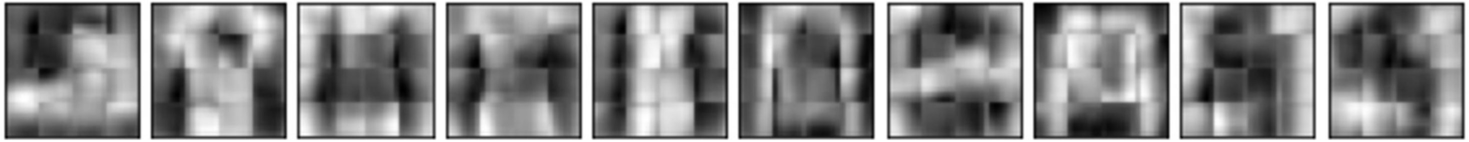
10 Images/160 Patches

15 Runs on Each Device for Each Patch – Best Reconstruction Saved

Original

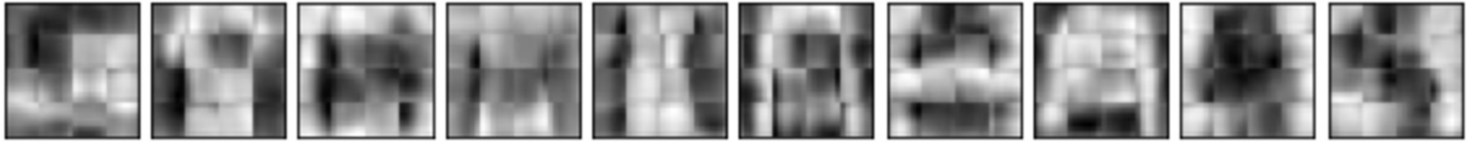


Loihi CSP



Average Reconstruction Error 2-Norm Loihi: .453 (std .13)
Average Number of Active Features Loihi: 5.0313 (std 1.05)

D-Wave



Average Reconstruction Error 2-Norm D-Wave: .471 (std .15)
Average Number of Active Features D-Wave: 5.856 (std 1.46)

MAIN TAKEAWAY: Loihi shows slightly better reconstructions (NOT statistically significant : Welch’s t-test pvalue = .24)
Loihi uses fewer coefficients (YES statistically significant : Welch’s t-test pvalue = 1.62 e-08)

Key take home messages

1. Bio-inspired sparse coding allows for unsupervised learning and optimization on neuromorphic devices at a fraction of the power and energy of other devices.
2. Loihi post-hoc normalization gives 50x faster solutions than regular Loihi with similar coefficient activity.
3. D-Wave is solving binary sparse coding in the form of QUBO.
4. Loihi can also perform binary sparse coding on much larger problems than D-Wave while also having flexibility to and solve for rate coded solutions.
5. Next Steps: a) Compare QUBO power and energy performance
   b) Compare D-Wave and Loihi on other QUBO problems
   c) Approximate QUBO with spiking LCA by shrinking time constant

Kyle Henke // Student
Benjamin Migliori // Neuromorphic Lead
Garrett Kenyon // Neuromorphic mentor
Nga Nguyen // D-Wave mentor

Thank you!

Further questions or interest in our code or results?

khenke@lanl.gov
505.330.2401

National Nuclear Security Administration

Los Alamos
NATIONAL LABORATORY
EST. 1943