Lessons from Loihi for the Future of Neuromorphic Computing

Mike Davies, Director of Intel’s Neuromorphic Computing Lab
March 16, 2021

Neuro-Inspired Computing Elements (NICE) 2021
Consider autonomous drone racing
Brains remain unrivaled computing devices

**Cockatiel Parrot**
- Brain Power: 50 mW
- Mass: 2.2 grams
- Can learn to speak English words
- Navigates and learns unknown environments at 35 km/h
- Can learn to manipulate cups for drinking

**Autonomous Drone**
- CPU/GPU controller
  - Power: 18,000 mW
  - Mass: ~40 grams
  - Pre-trained to fly between known gates at walking pace
  - Can't learn anything online

Deep learning models are increasingly power hungry

Two Distinct Eras of Compute Usage in Training AI Systems
Petaflop/s-days

300,000x increase in required training computation over 6 years ... versus 8x provided by Moore’s Law

Not on a trajectory to close the efficiency gap with nature!

Source: OpenAI [https://openai.com/blog/ai-and-compute/]
Deep learning is fundamentally limited in other respects

<table>
<thead>
<tr>
<th>Natural Learning</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Fast generalization with few examples</td>
<td>▪ Slow generalization with massive data</td>
</tr>
<tr>
<td>▪ Online and incremental</td>
<td>▪ Offline and batched</td>
</tr>
<tr>
<td>▪ Automatic abstraction</td>
<td>▪ “Curve fitting”</td>
</tr>
</tbody>
</table>
Our Approach: Look to the brain, co-design the architecture and algorithms

**Novel Neuro-Inspired Algorithms**

<table>
<thead>
<tr>
<th>Category</th>
<th>Example applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning: backprop-trained event-based DNNs</td>
<td>Object and gesture recognition for event-based vision sensors, slip detection for event-based tactile sensors, ANNs with sparsely changing input data</td>
</tr>
<tr>
<td>Deep learning: DNNs with online adaptation</td>
<td>Few-shot new gesture learning, Adaptive control,</td>
</tr>
<tr>
<td>Vector Symbolic Architectures (VSA), aka Hyperdimensional Computing (HDC)</td>
<td>Semantic factorization, relational reasoning, symbolic and analogical reasoning</td>
</tr>
<tr>
<td>Neural Engineering Framework (NEF)</td>
<td>Adaptive control systems, state machines</td>
</tr>
<tr>
<td>Dynamic Neural Fields (DNF)</td>
<td>SLAM, object tracking, dynamic control, attention</td>
</tr>
<tr>
<td>Neural sampling e.g. spiking Boltzmann machines</td>
<td>Constraint satisfaction, probabilistic inference</td>
</tr>
<tr>
<td>Oscillatory computation</td>
<td>Optimization, event-based spectral transforms, optic flow, audio spectral normalization</td>
</tr>
<tr>
<td>Recurrent Excitation/Inhibition-balanced networks</td>
<td>LASSO regression, sparse feature coding</td>
</tr>
<tr>
<td>Event-based networks with temporally coded information</td>
<td>Graph search, similarity search</td>
</tr>
</tbody>
</table>

**Neuro-Inspired Silicon**

- **Pursue neuro inspiration?**
  - Yes
  - No

- **Compute-memory integrated?**
  - Yes
  - No

- **Temporal neuron models?**
  - Yes
  - No

- **Standard CMOS or new devices?**
  - New
  - CMOS

- **Asynchronous design style?**
  - Yes
  - No

- **Integrate analog circuits?**
  - Yes
  - No

- **Support plasticity? (+ other novel features)**
  - Yes
  - No

### Conventional Deep Networks

\[ u_i = \sum_j w_{ij} f(u_j) + b_i \]

### Neuromorphic Networks

\[ u_i(t) = \sum_j w_{ij} (\delta_j(t) * a_u(t)) + b_i \]

\[ r\psi(t) = (-v_i(t) + u_i(t)) - V_{thr}\delta_i(t) \]
Motivates a fundamentally different kind of computing

Parallel Computing

Batched + Vectorized Processing

Vector-valued Functions

$f(x)$ learned with backprop

Input

Output

DATA

Unbatched + Sparse Processing

Event-Driven Dynamical Systems

Function self-organizes

Neuromorphic Computing
Our **Loihi** chip

**KEY PROPERTIES**

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute and memory integrated</td>
<td>to spatially embody programmed networks</td>
</tr>
<tr>
<td>Temporal neuron models (LIF)</td>
<td>to exploit temporal correlation</td>
</tr>
<tr>
<td>Spike-based communication</td>
<td>to exploit temporal sparsity</td>
</tr>
<tr>
<td>Sparse connectivity</td>
<td>for efficient dataflow and scalability</td>
</tr>
<tr>
<td>On-chip learning</td>
<td>without weight movement or data storage</td>
</tr>
<tr>
<td>Digital asynchronous implementation</td>
<td>for power efficiency, scalability, and fast prototyping</td>
</tr>
</tbody>
</table>

Yet...

<table>
<thead>
<tr>
<th>Limitation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No floating-point numbers</td>
<td>No multiply-accumulators</td>
</tr>
<tr>
<td>No off-chip DRAM</td>
<td></td>
</tr>
</tbody>
</table>

Fundamental to deep learning hardware
Intel Neuromorphic Research Community

Collaborating to Accelerate the Research

INRC includes over 120 groups

Other names and brands may be claimed as the property of others.
Loihi Results
**Efficient Sensing**

**Gesture recognition + learning**
Loihi + DAVIS 240C camera
60 mW total power, 15 mW dynamic [Task 5]
G. Orchard and SB Shrestha, with K. Stewart, E. Neftci (UCI)

**Visual-Tactile Sensing**
45x lower power
20% faster vs GPU [Task 6]
T. Taunyazov et al (NUS)

**Olfaction-inspired odor recognition and learning**
3000x more data efficient learning than a deep autoencoder
Nabil Imam and Thomas Cleland, Nature Machine Intelligence, March 2020

**Audio keyword spotting**
>100x lower energy per inference vs GPU [Task 1]
P. Blouw et al (ABR)

See backup for references and configuration details. Results may vary.
Compelling results for robotic and drone workloads

Adaptive robotic arm control
40x lower power, 50% faster vs GPU
Task 8
Applied Brain Research

iCub scene understanding
Integrated behaviors: Object recognition, tracking, learning
with A. Glover, C. Bartolozzi (IIT)

Event-based UAV horizon tracking
DVS Hough transform +
Adaptive PID controller
2ms latency,
22x faster vs CPU
Task 14
Intel/ETHz

Head Direction
Localization and Learning
100x lower power vs CPU
Task 10
G. Tang, K. Michmizos (Rutgers)
Y. Sandamirskaya et al (Intel/ETHz/NI)

Micro Aerial Vehicle Landing
Evolutionary design of a 35-neuron network that achieves smooth MAV landings with Loihi on board

See backup for references and configuration details. Results may vary.
Even greater gains for sparse computational studies

Graph Search
With temporally coded spike wavefronts
100x faster vs CPU [Task 12]

Combinatorial optimization
(CSP, SAT, sudoku, train scheduling)
2,000x lower energy and 40x faster vs CPU [Task 13]

Heat diffusion modeling
Scaled to 100+ chips and 300k mesh points
B. Aimee et al (Sandia)

LASSO / sparse reconstruction
( Locally Competitive Algorithm)
10^3x faster, 10^4x lower energy vs CPU [Task 9]

Hear more at the Loihi tutorial!

See backup for references and configuration details. Results may vary.
For the Right Workloads, Loihi Provides Orders of Magnitude Gains in Latency and Energy

Reference architecture:
- CPU (Intel Core/Xeon)
- GPU (Nvidia)
- Movidius (NCS)
- TrueNorth

See backup for references and configuration details. Results may vary.
Standard feed-forward deep neural networks give the least compelling gains (if gains at all)

See backup for references and configuration details. Results may vary.
Recurrent networks with novel bio-inspired properties give the best gains.
Compelling scaling trends: Larger networks give greater gains

See backup for references and configuration details. Results may vary.
Deep Learning on Loihi

Red: ANNs converted with rate coding
- Low energy but high latency
- Poor scaling
- Not very promising

Blue + purple: Offline backprop-trained spike timing
- Low energy and low latency
- Compute intensive to train (and scale)

Green: Online backprop
- Well suited for continuous adaptation

See backup for references and configuration details. Results may vary.
Loihi shows order of magnitude gains are possible

- In energy efficiency
- In speed of processing data – especially signals arriving in real time
- In the data efficiency of learning and adaptation
- With programmability to span a wide range of workloads and scales
- Long term, we will need to reduce cost with process technology innovations
Computing with Collective Dynamics

**Gradient Descent**
- Backprop (offline)
- Online Backprop approximations

**Non-Gradient Based Approaches**
- Olfaction-inspired learning
- Associative learning (e.g. SLAM)
- Graph Search

**Plastic Weights**
- Locally Competitive Algorithm
- Winner Take All
- Dynamic Neural Fields

**Static Weights**
- Combinatorial optimization
- Nearest Neighbor Search
Neuromorphic Learning Perspectives

Gradient-Based Learning
- DNN scaling possible (?), not yet proven
- Data hungry – slow to learn
- Data samples need to be uniformly distributed during learning
- Learning activity is not sparse

Limited today to shallow networks that run relatively slowly
Examples: feedback alignment, e-prop, delta
Good for fine-tuning and adapting

Non-Gradient Based Learning
- No “deep” examples to date
- Fast to learn from few examples
- Networks mostly need to be hand engineered and tuned
- Learning activity is sparse

Limited today to interesting examples, but with narrow scope
Example: olfactory model
Good for associative learning
Outlook to Commercialization

- Control
- Assistive robotics manufacturing
- Delivery
  - Structural inspections
  - Agriculture, Mining
- Event-based vision
- Surveillance, Security
- Smart sensors
- In car
- Outside car
- Infotainment
- Safety
- Autonomy

- Speech recognition
- Noise suppression
- Gesture recognition
- Human-computer interfaces
- AR/VR
- Resource scheduling
- Anomaly detection
- Recommendation systems

SDK/SW Tools
Design Services
System OEMs
Supplier Ecosystem
Developer Community

Neuromorphic Computing Lab
intel labs
Legal Information

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Results have been estimated or simulated.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.
Thank You!

Learn more at the Loihi tutorial tomorrow
References and System Test Configuration Details

[Task 4] Loihi: Nahuku system running NxSDK 0.95. CIFAR-10 image recognition network trained using the SNIN-Toolbox (code available at https://snntoolbox.readthedocs.io/en/latest/). CPU: Core i7-9700K with 32GB RAM, GPU: Nvidia RTX 2070 with 8GB RAM. OS: Ubuntu 16.04.6 LTS, Python: 3.5.5, TensorFlow 1.131. Performance results are based on testing as of July 2020 and may not reflect all publicly available security updates.