Braindrop: A Mixed-Signal Neuromorphic System that Presents Clean Abstractions

Kwabena Boahen*
Bioengineering & Electrical Engineering
Stanford University

*Cofounder & Chief Scientific Adviser, Femtosense Inc.

26 March 2019
Deep learning is huge — in the cloud

- Backprop learning is powerful
- Networks deep in space or time
- Space is discretized into layers
- Time is discretized into steps
- Unit’s output must be differentiable (with respect to outputs of units feeding it)
Backprop’s constraints limit design-space

- Cannot take advantage of:
  - Physical **space** (its continuous)
  - Real **time** (its also continuous)
  - Non-differentiable signals (e.g., **spikes**)

- ✤
How do we relax its constraints? (Part I)

- Map functional abstractions onto physical ones
- Two existing examples:
  - Neural Engineering Framework (Eliasmith & Anderson 2003)
  - Predictive Coding Framework (Deneve et al. 2014)
How do we relax its constraints? (Part II)

- Train networks continuous in time and space
- Known as dynamical systems
- An existing example:
  - Neural Ordinary Differential Equations (Duvenaud et al. 2018)

![Diagram showing Residual Network and ODE Network](image)
Exploit physical primitives to implement physical abstractions

Reap dramatic gains in energy-efficiency

- Harvest Vibration: 500 µW
- Sense Accelerometer: 6 µW
- Compute Neuroprocessor: 100 µW
- Communicate BTLE (1% on-time): 280 µW
Minimizing energy

\[ E_{op} = N_{active}N_{conn} = (\rho_{active}N)(\rho_{conn}N) \]
Minimizing energy: Temporal sparsity

\[ E_{\text{op}} = N_{\text{active}} N_{\text{conn}} = (\rho_{\text{active}} N)(\rho_{\text{conn}} N) \]
Temporal sparsity: Spikes

\[ E_{op} = N_{active} N_{conn} = (\rho_{active} N)(\rho_{conn} N) \]
Minimizing energy: Spatial sparsity

\[ E_{op} = N_{active} N_{conn} = (\rho_{active} N) (\rho_{conn} N) \]
Spatial sparsity: Analog convolving

\[ E_{op} = N_{active} N_{conn} = (\rho_{active} N) (\rho_{conn} N) \]
Make point about top-left corner being lowest power
Digital versus Analog: 1 day versus 1000 yrs

4.3B-transistor processor
10.35Wh battery

28 nm FDSOI thick-oxide transistor

2.5 hours
1 day
10 days
100 days
2.74 years
27.4 years
274 years
2747 years
Analog Challenge I: Heterogeneity
Silicon neurons’ tuning-curves (Braindrop)

Top: Fig. 2. Silicon neurons’ tuning-curves (Braindrop)

Since tuning curves are measured in the presence of random noise, the weight applied to the neuronal responses (see Fig. 1). The approximation decoded information from the input signal into time-varying spike-rates across the population at a single temperature (26°C).

484 Braindrop neurons at 26°C
Analog Challenge I: Thermal Sensitivity

\[ I_\mu (T) = I_{0_{\text{nom}}} e^{(\gamma_1)(1 - \frac{T_{\text{nom}}}{T}) e^{\frac{(1 - \kappa)V_{BS}}{U_T}}} \]
\[ \times e^{\frac{\kappa V_{GS}}{U_T} e^{(\lambda_1 T_{\text{nom}} + \lambda_2)\Delta V_{DS}} \left(1 - e^{-\frac{V_{DS}}{U_T}}\right)} \]

- A subthreshold transistor’s current \((I_\mu)\) is exponentially sensitive to temperature
- \(T\) is the absolute temperature
- \(U_T = kT/q\) is the thermal voltage
- Across a 50°C range, the current changes by 1.5 to 3 decades

\[ I_{0_{\text{nom}}} = 0.2 \mu A, \quad \gamma_1 = 10, \quad \kappa = 0.6, \quad \lambda_1 = 0.005, \quad \lambda_2 = 0.001, \quad \Delta V_{DS} = 500 \text{ mV} \]

\[ V_{GS} = 0.1 \text{ V} \]

\[ V_{DS} = 0 \text{ V} \]
Tuning-curves’ thermal sensitivity (Braindrop)

Reid, Montoya, & Boahen 2019

4 Braindrop neurons for 0 to 38°C
Approximating functions

- The desired function $f(x)$ is expressed as a weighted sum of the neural tuning curves $a_i(x)$.
- The weights—called decoders—are labeled $d_i$.

$$f = A d \Rightarrow d = (A^T A)^{-1} A^T f$$

3 to 30 Braindrop neurons at 26°C
Thermally robust computation (Braindrop)

256 Braindrop neurons from 0 to 38°C
NEF: Decode-Transform-Encode

1. Spike
\[ \delta_{x_i} = [0, \alpha_i(I_i + \beta_i)] \]
Somas emit unit-area deltas \( \delta_{x_i} \) at rates \( \langle \delta_{x_i} \rangle_t \) dictated by their input current \( I_i \).

2. Decode
\[ \delta_{y_i} = \sum_j D_{ji} \delta_{x_i}, D \in \mathbb{R}^{D \times N} \]
Deltas are then scaled by their decode weight and merged together.

3. Transform
\[ \delta_{z_k} = \sum_j T_{kj} \delta_{y_i}, T \in \mathbb{R}^{D \times D} \]
Transform works the same way as Decode.

4. Encode
\[ \tau \dot{I}_i = -I_i + \sum_k E_{ik} \delta_{z_k}, E \in \mathbb{R}^{N \times D} \]
Synaptic filters superpose and low-pass filter weighted deltas to produce output currents \( \langle I_i \rangle \) that feed the next soma layer.

Eliasmith & Anderson 2003
Digital thinning and analog convolving

1. **Spike**
   \[ \langle \delta_{x_i} \rangle_t = [0, \alpha_i(I_i + \beta_i)] \]
   Somas emit delta trains (as in Figure 2).

2. **Decode**
   \[ \langle \delta_{y_i} \rangle_t = \sum_i D_{ji} \langle \delta_{x_i} \rangle_t \]
   \[ D \in [-1, 1]^{D \times N} \]
   Weighted deltas are accumulated to produce a stream of unit-area deltas.

3. **Transform**
   \[ \langle \delta_{z_k} \rangle_t = \sum_j T_{kj} \langle \delta_{y_i} \rangle_t \]
   \[ T \in [-1, 1]^{D \times D} \]
   Transform still works the same as Decode (\( T_{kj} = 1 \) in this example).

4. **Sparse Encode**
   \[ \tau I_l = -I_l + \sum_k S_{lk} \delta_{z_k} \]
   \[ S \in \{-1, 0, 1\}^{N \times D} \]
   Each accumulator’s deltas are sent to a subset of synaptic filters (tap-points).

5. **Convolve**
   \[ I_m = \sum_t \hat{d}_\gamma (m - t) I_l \]
   Filter outputs (\( I_l \)) are convolved (\( I_m \)) and sent to the next soma layer.
Somas’ spikes are muxed into address-events.

Address-events are translated into base-addresses.

Decoding vector and buckets’ states are retrieved and the latter are updated. A tag is emitted if threshold is exceeded.

Tags are buffered.

Tags are translated into base-addresses.

Address-events are demuxed to excite or inhibit synaptic filters.

Filters’ output currents are distributed to somas.

Corresponding transform’s column and associated buckets’ states are retrieved, states updated, and tags emitted.

Tags are translated into address-events with associated polarities.
Braindrop: 4096 neurons in 28nm FDSOI CMOS
**Programming Environment (Nengo)**

\[ a = f(x) \]
\[ b = g(y) \]
\[ \dot{z} = h(z) + a + b \]
To emulate the dynamical system

\[ \tau_{\text{dyn}} \dot{x}(t) = f(x) + u(t) \]

Choose decoding weights such that, after synaptic filtering,

\[ \sum_i D_i^T \delta x_i \approx \frac{\tau_{\text{syn}}}{\tau_{\text{dyn}}} f(x) + x \]

And set

\[ c(t) = \frac{\tau_{\text{syn}}}{\tau_{\text{dyn}}} u(t) \]

Eliasmith & Anderson 2003
Tapped delay-line (Braindrop)

\[
\theta \dot{x}(t) = Ax(t) + Bc(t)
\]

\[
c(t - \theta') \approx C_{\theta' / \theta} x(t), \quad 0 \leq \theta' \leq \theta
\]
**Measured Energy/op (pJ)**

1. Somas’ spikes are muxed into address-events.
2. Address-events are translated into base-addresses.
3. Decoding vector and buckets’ states are retrieved and the latter are updated. A tag is emitted if threshold is exceeded.
4. Tags are buffered.
5. Tags are translated into base-addresses.
6. Corresponding transform’s column and associated buckets’ states are retrieved, states updated, and tags emitted.
7. Tags are translated into address-events with associated polarities.
8. Address-events are demuxed to excite or inhibit synaptic filters.
9. Filters’ output currents are distributed to somas.

- **1.1**: 3.8
- **15**: 15
- **28**: 6.4
- **3**: 1.1
* Analog convolving fans out $d$ spike-trains to $N$ neurons; sparsifies spatially by $d/N$

* Digital thinning lets one per SNR spikes through; sparsifies temporally by $1/\text{SNR}$
Task performance

- Two components:
  - Network design
  - Hardware design

\[ E_{sys}(task) = E_{HW}(R_{net}(task)) \]

\[ E_{op} = N_{active} N_{conn} = (\rho_{active} N)(\rho_{conn} N) \]
Acknowledgments

Collaborators
Stanford
Krishna Shenoy
Tirin Moore
Oussama Khatib
Waterloo
Chris Eliasmith
Terry Stewart
Aaron Voelker
Cornell & Yale
Rajit Manohar
Ned Bingham
Silvia Ferrari
Taylor Clawson

Students & Staff
Max Kanwal
Scott Reid
E. Kauderer-Abrams
Jonathan Timcheck
Recent Alumni
Ben Benjamin
Alex Neckar
Sam Fok
Nick Oza
Ashok Cutkosky
John Aguayo
Tatiana Engel
Samir Menon
Peiran Gao
Nick Steinmetz
Rodrigo Alvarez

Funding
NIH
Pioneer,TR01
ONR: C. Baatar
Braindrop: A Mixed-Signal Neuromorphic Architecture With a Dynamical Systems-Based Programming Model

This paper provides an overview of a current approach for the construction of a programmable computing machine inspired by the human brain.

By Alexander Neckar®, Sam Fok®, Ben V. Benjamin, Terrence C. Stewart, Nick N. Oza, Aaron R. Voelker®, Chris Eliasmith®, Rajit Manohar®, Senior Member IEEE, and Kwabena Boahen, Fellow IEEE