Lecture: Spiking Networks Intro

- Topics: intro to spiking neurons, coding, training
Biological Neuron

Source: stackexchange.com
Biological Neuron

- Input = dendrite; output = axon; connection = synapse
- Neurotransmitters flow across the synaptic cleft; the receiving (post-synaptic) neuron opens up ion channels, allowing extracellular ions to flow in and impact membrane potential
- Synapses can be excitatory or inhibitory
- The neuron fires when its potential reaches a threshold; output spike lasts 1-2ms and is about 100mV
- Resting potential is -65mV; the threshold is typically 20-30mV higher; each input spike raises the potential by about 1mV; need 20-50 input spikes to trigger an output spike
- After an output spike, the potential falls below the resting potential
- After an output spike, channels remain open for a refractory period; during this time, input spikes have a very small impact on potential
- A single neuron may connect to >10K other neurons; ~100 billion neurons in human brains; ~500 trillion synapses
Neuron Models

• Hodgkin and Huxley took many measurements on a giant axon of a squid; that led to differential equations to quantify how ion flow impacts neuron potential

• Several experiments since then to better understand neuronal behavior, e.g., there are different kinds of neurons that respond differently to the same stimulus
Red Herring

• Some detail is good, e.g., some neurons have adapted themselves in a certain way because it’s computationally more powerful (the refractory period, neurons that act as integrators vs. neurons that look for a resonant input frequency)

• Some detail is unhelpful, e.g., complicated equations that capture the non-idealities of bio-chemical processes (the refractory period, the exact shape of the potential curve)
The LIF Model

Also note Linear LIF

Source: Nere et al., HPCA’13
Neuronal Codes

Source: Kruger and Aiple, 1988
Rate Vs. Temporal Codes

• Rate code: information is carried in terms of the frequency of spikes ➔ relative timing of spikes on two dendrites is irrelevant

• Temporal code: information is carried in terms of the exact time of a spike (time to first spike or a phase code)

• There are experiments that have observed both codes

• The same code can apply throughout a multi-layer network; the code can incorporate stochastic elements; a new input is presented after a “phase” expires
Rate Codes

- The most popular approach today
- Output freq = w1*input freq1 + w2*input freq2
- Works fine with integrator neurons; will also work fine with neurons that look for resonant frequencies
- Needs multiple spikes to encode a value (more energy)
Temporal (Spike) Codes

- Works fine with integrator neurons
- What separates two different input values is the intervening leak
- Identifies the tail end of a weighted cluster of inputs (also a useful way to find correlations, but it’s not the neural network equation we know)
- Needs a single spike to encode a value (less energy)
Rate Vs. Temporal Codes

Temporal Codes

(a) Single red spike

(b) Red and pink spikes

Time

50 125

In 1

In 2

Potential

Output
Rate Vs. Temporal Codes

Temporal Codes

(c) Blue and cyan spikes

(d) Red and blue spikes
Rate Vs. Temporal Codes

Temporal Codes
(d) Red and blue spikes

Rate Codes
(e) Red and blue spikes
Training

• Can use back-prop and supervised learning (has yielded the best results so far)

• More biologically feasible: Spike Timing Dependent Plasticity (STDP)

• If a presynaptic spike triggers a postsynaptic spike, strengthen that synapse

• This is an example of unsupervised learning; the network is being trained to recognize certain clusters of inputs
STDP

Pre before post

Source: Scholarpedia.org
The Spiking Approach

- Low energy for computation: only adds, no multiplies
- Low energy for communication: depends on spikes per signal
- Neurons have state, inputs arrive asynchronously, info in relative timing of spikes, other biological phenomena, ...
IBM TrueNorth

• Product of DARPA’s SyNAPSE project

• Largest chip made by IBM (5.4 billion transistors)

• Based on LLIF neuron model

• Lots of on-going projects that use TrueNorth to execute new apps

• Lots of limitations as well – all done purposely to reduce area/power/complexity
TrueNorth Core

Input spikes

Axons

Crossbar Synapses

Type

Axon Activity $A(t)$

Neurons

Select & Encode

Output spikes

Sync

$A_1, A_3$
TrueNorth Core (Axonal Approach)

### Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_{ji}$</td>
<td>Connection between axon $j$ and neuron $i$</td>
<td>0, 1</td>
</tr>
<tr>
<td>$G_j$</td>
<td>Axon type</td>
<td>0, 1, 2</td>
</tr>
<tr>
<td>$S_i^{0..2}$</td>
<td>Synapse values</td>
<td>-256 to 255</td>
</tr>
<tr>
<td>$L_i$</td>
<td>Leak</td>
<td>-256 to 255</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Threshold</td>
<td>1 to 256</td>
</tr>
</tbody>
</table>
TrueNorth Core (Dendritic Approach)
Principles

• Focus on low power (speed is secondary)

• Low power with asynchronous circuits

• Low power with high-Vt circuits and SOI process

• A dendritic approach leads to fewer SRAM reads/updates and is more deterministic

• Also uses an under-provisioned network (giving rise to the concept of “future” ticks)
Receiving Spikes

• Spikes arrive at the *Scheduler*; stored in a 256x16 SRAM grid to indicate axon and time of the spike

• The *Token Controller* receives spikes in a “*tick*” (and their types – exc/inh); it then sequentially walks through 256 neurons – this dendritic approach makes the latency and SRAM accesses more deterministic

• It reads a 410-bit word from SRAM for that neuron (this word has connectivity info used by Token Controller and neuron parameters used by *Neuron block*

• Based on connectivity and input spikes, instructions are sent to Neuron block
Neuron Block

- Neuron computations are performed here (these are non-trivial because of stochastic behaviors – more later)

- A leak is introduced every tick for every neuron

- After thresholding, a spike may be triggered

- The neuron block uses synchronous circuits, but it is active only when it receives instructions from the Token Controller; the Token Controller, Scheduler, and network Router are all asynchronous to exploit the low power inherent in spike sparsity
More Details

• The 410 bit word: 256 synaptic connections, 4 9-bit weights, 9-bit leak, 8-bit threshold, 26-bit destination axon, 4-bit delivery time, 9-bit potential, ... (62 more bits)

• The network may occasionally not handle a burst of spikes; if this is likely, then spikes must have a future delivery time; allows a spike to take multiple ticks to reach its destination

• 256x256 core; 4K cores; million neurons; 430 mm2; 65 mW on average; 28nm Samsung process; 1ms tick
Limitations

• A neuron can only have 256 inputs (no provision for partial sums)

• A neuron can only have one destination axon output of a specific type (seen by 256 neurons)

• Weight quantization (4 weights per neuron)
References

• “Cognitive Computing Building Block”, A. Cassidy et al., IJCNN’13

• “A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45 pJ per Spike in 45nm”, P. Merolla et al., CICC, 2011


• “Real-Time Scalable Cortical Computing…”, A. Cassidy et al., SC’14