Analog Computers for Deep Machine Learning

Jeremy Holleman
CTO, Syntiant
Assoc. Professor, UNCC
Always-On Machine Learning for Battery Powered Devices

**Market**
- ~$80B processor market for battery powered devices

**Applications**
- Mobile Phones
- Earbuds
- Smart Watches
- Drones
- Cameras

**Technology**
- Analog neural network
- Eliminates data movement penalty
- 50x lower power vs GPUs

**Status**
- Working silicon
- First commercial agreement

**Team**
- Median exp: 20 years
- 70% PhD

**Contact**
- Kurt Busch, CEO
- kurt@syntiant.com
Basic Thesis

Analog computation is efficient
Analog computation keeps information local
Memory locality is key
Neural networks are robust
Fundamental Efficiency of Analog Compute

Resolution (SNR) dictates local power consumption

Modest precision arithmetic can be done with extreme efficiency

\[ i_{n,RMS} = \sqrt{2qI_D B (A^2 / Hz)} \]

\[ E = T I_D V_{DD} = 2q2^{2N_{bits}} V_{DD} \]

\(~20 \text{ fJ } @ \text{ 8b (100 TOPS/W)}\)

\(~20 \text{ aJ } @ \text{ 3b (100 POPS/W)}\)
Memory access energy dominates most digital architectures

Analog compute-in-memory stores the weight in the multiplier device

Data flow architectures avoid round-trips to cache
  - Ex: Fetch from 32kB SRAM (45nm) ~ 20pJ*

Density of analog compute allows high parallelism and keeps operators physically close

The Immunity Advantage of Digital

Digital logic works so well because errors don’t propagate

But signal restoration is not limited to binary signals

A hard problem can be decomposed into many easier problems

\[
\begin{array}{c}
1101 \\
\times 1001 \\
\hline
1101 \\
0000 \\
0000 \\
+ 1101 \\
\hline
1110101
\end{array}
\]

Partial Products
- 4 4b × 1b PPs
- OR
- 16 1b × 1b PPs

Discretization between stages
The Immunity Advantage of Digital

A hard problem can be decomposed into many easier problems
With noise rejection in between stages

Explicitly, as in multiplication

OR Implicitly, as in neural networks

Discretization between stages
K-Means Clustering Algorithm

- Online K-Means
  - Initialize centroids
  - Assign points
  - Move centroids toward points

\[
\tilde{C}_{i+1} = \begin{cases} 
(1 - \lambda)\tilde{C}_i + \lambda\tilde{x} & \text{if } x \in C \\
C_i & \text{else}
\end{cases}
\]
I-D Conditional mean and variance learner is the core computational unit.

- AAE: analog arithmetic elements
- FGM: floating gate
Transistor Size Scaling

- Robustness to static error inherent in learning algorithm allows aggressive device size reduction
- System modeling and simulation provides knowledge of the system’s tolerance to mismatch errors

Young, et al. TNNLS 2013
Clustering Results

- Input Data
- Cluster Means
- Evolution of Centroid Means
- Extracted Variance

\[ \mu = \sigma^2 = D \]

\[ y(n_A) \]

\[ \mu = (3, 3) \]
\[ \sigma^2 = (1, 1) \]

\[ \mu = (7, 7) \]
\[ \sigma^2 = (1.5, 1.5) \]

\[ \mu = (7, 3) \]
\[ \sigma^2 = (1, 1) \]

\[ \mu = (3, 7) \]
\[ \sigma^2 = (0.6, 0.6) \]
Comparison

Synthesized equivalent digital circuit for comparison
Analog Implementation 288x more efficient

<table>
<thead>
<tr>
<th>Energy per Evaluation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dig.</td>
<td>7.7 nJ</td>
</tr>
<tr>
<td>Ana.</td>
<td>27 pJ</td>
</tr>
</tbody>
</table>

Lu, et al. ISSCC 2014
Conclusions

- Question assumptions
- Find out how good is good enough
- Machine learning is highly empirical for now. Test everything
- Bring experts from different fields together
  - Learn to interact with people outside your field
Thank You
Syntiant Analog Compute-in-Flash

Hardware Demonstration

Currents scaled by programmable threshold provides NN weight. Exploits sub-threshold exponential voltage-current relationship

Core Technology

Performs the matrix-vector operations central to all major neural networks (CNN, LSTM, …) with ultra-low-power and fully parallel computation.

Illustrates a weighted sum of two input waveforms with arbitrary programmable weight ratio.

\[ I_{Out,Sum} = w_1I_{In1} + w_2I_{In2} \]

\[ I_{Out} = I_{In} e^{\frac{V_{Th,In} - V_{Th,Out}}{nUT}} \]

\[ I_{Out,j} = \sum_{i=0}^{N} I_{In,i}w_{i,j} \]
Syntiant Analog Compute-in-Flash

- First step is to program eFLASH devices with different weights
  - Changes the slope of sub-threshold region
- Sets the Neural Network coefficients and ability to perform power-efficient Multiple Accumulate (MAC) over many devices

2 Devices w/ Different Weights
• Inject unique inputs into the memory to demonstrate Multiply and Accumulate
  ○ Push 2 sine waves into two separate devices into the eFLASH memory (1Hz and 2.5 Hz)
  ○ Waveforms are scaled, summed in the network with different **Weights**
Thinking Differently

Conventional
Known procedure
generates answers

Deep Learning
Known answers
generate procedures
with training

Quantum
Answers superimposed
Select and measure the answer
The Future Third Wave of AI

Expert Systems
“Symbolist Prevail”

Handcrafted Knowledge
- Perceiving
- Learning
- Abstraction
- Reasoning

Deep Learning
Connectionist Dominate

Statistical Learning
- Perceiving
- Learning
- Abstraction
- Reasoning

Third Wave
Contextual Adaption
- Perceiving
- Learning
- Abstraction
- Reasoning

Taken from DARPA Perspective on AI by John Launchbury