Rethinking computation: A processor architecture for machine intelligence

Amir Khosrowshahi
Co-founder and CTO, Nervana
About nervana

- A platform for machine intelligence
- enable deep learning at scale
- optimized from algorithms to silicon
Model and substrate for computation

Mammalian cortex  Functional model  Machine learning model

Hard!
Model and substrate for computation

Do this instead:

Custom ASIC

- Model description language
- Hardware abstraction layer
- Distributed primitives
- Compilers, drivers

Feasible, but still hard.

Deep learning model
Application areas

- Healthcare
- Agriculture
- Finance
- Online Services
- Automotive
- Energy
nervana cloud

Data

Images
Video
Text
Speech
Tabular
Time series

Cloud

import
build
train
deploy
Deep learning as a core technology

'Google Brain' model

Nervana Platform
- Fastest library

<table>
<thead>
<tr>
<th>Library</th>
<th>Class</th>
<th>Time (ms)</th>
<th>forward (ms)</th>
<th>backward (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nervana-neon-fp16</td>
<td>ConvLayer</td>
<td>230</td>
<td>72</td>
<td>157</td>
</tr>
<tr>
<td>Nervana-neon-fp32</td>
<td>ConvLayer</td>
<td>270</td>
<td>84</td>
<td>186</td>
</tr>
<tr>
<td>CuDNN[R4]-fp16 (Torch)</td>
<td>cudnn.SpatialConvolution</td>
<td>462</td>
<td>112</td>
<td>349</td>
</tr>
<tr>
<td>CuDNN[R4]-fp32 (Torch)</td>
<td>cudnn.SpatialConvolution</td>
<td>470</td>
<td>130</td>
<td>340</td>
</tr>
<tr>
<td>Chainer</td>
<td>Convolution2D</td>
<td>687</td>
<td>189</td>
<td>497</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>conv2d</td>
<td>905</td>
<td>187</td>
<td>718</td>
</tr>
<tr>
<td>Caffe</td>
<td>ConvolutionLayer</td>
<td>1935</td>
<td>786</td>
<td>1148</td>
</tr>
<tr>
<td>CL-nn (Torch)</td>
<td>SpatialConvolutionMM</td>
<td>7016</td>
<td>3027</td>
<td>3988</td>
</tr>
<tr>
<td>Caffe-CLGreenTea</td>
<td>ConvolutionLayer</td>
<td>9462</td>
<td>746</td>
<td>8716</td>
</tr>
</tbody>
</table>
Fastest library

<table>
<thead>
<tr>
<th>Library</th>
<th>Class</th>
<th>Time (ms)</th>
<th>forward (ms)</th>
<th>backward (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nervana-neon-fp16</td>
<td>ConvLayer</td>
<td>230</td>
<td>72</td>
<td>157</td>
</tr>
<tr>
<td>Nervana-neon-fp32</td>
<td>ConvLayer</td>
<td>270</td>
<td>84</td>
<td>186</td>
</tr>
<tr>
<td>NVIDIA / Facebook</td>
<td>cudnn.SpatialConvolution</td>
<td>462</td>
<td>112</td>
<td>349</td>
</tr>
<tr>
<td>NVIDIA / Facebook</td>
<td>cudnn.SpatialConvolution</td>
<td>470</td>
<td>130</td>
<td>340</td>
</tr>
<tr>
<td>Chainer</td>
<td>Convolution2D</td>
<td>687</td>
<td>189</td>
<td>497</td>
</tr>
<tr>
<td>Google</td>
<td>conv2d</td>
<td>905</td>
<td>187</td>
<td>718</td>
</tr>
<tr>
<td>Berkeley</td>
<td>ConvolutionLayer</td>
<td>1935</td>
<td>786</td>
<td>1148</td>
</tr>
<tr>
<td>CL-nn (Torch)</td>
<td>SpatialConvolutionMM</td>
<td>7016</td>
<td>3027</td>
<td>3988</td>
</tr>
<tr>
<td>Caffe-CLGreenTea</td>
<td>ConvolutionLayer</td>
<td>9462</td>
<td>746</td>
<td>8716</td>
</tr>
</tbody>
</table>
• Fastest library
• Model support

Models
• Convnet
• RNN, LSTM
• MLP
• DQN
• NTM

Domains
• Images
• Video
• Speech
• Text
• Time series
Running locally:

```bash
% python rnn.py  # or neon rnn.yaml
```

Running in nervana cloud:

```bash
% ncloud submit -py rnn.py  # or --yaml rnn.yaml
% ncloud show <model_id>
% ncloud list
% ncloud deploy <model_id>
% ncloud predict <model_id> <data>  # or use REST api
```
nervana neon

- Fastest library
- Model support
- Cloud integration
- Multiple backends

Backends
- CPU
- GPU
- Multiple GPUs
- Parameter server
- (Xeon Phi)
- nervana TPU
• Fastest library
• Model support
• Cloud integration
• Multiple backends
• Optimized at assembler level
nervana tensor processing unit (TPU)

- Unprecedented compute density
nervana tensor processing unit (TPU)

- Unprecedented compute density
- Scalable distributed architecture
nervana tensor processing unit (TPU)

- Unprecedented compute density
- Scalable distributed architecture
- Memory near computation
nervana tensor processing unit (TPU)

- Unprecedented compute density
- Scalable distributed architecture
- Memory near computation
- Learning and inference
- Exploit limited precision
- Incorporate latest advances
- Power efficiency
nervana tensor processing unit (TPU)

- Unprecedented compute density
- Scalable distributed architecture
- Memory near computation
- Learning and inference
- Exploit limited precision
- Incorporate latest advances
- Power efficiency

- 10-100x gain
- Architecture optimized for algorithm
General purpose computation

Motivation: reduce power and cost, fungible computing.

Enabled inexpensive mobile devices.

2000s: SoC
Dennard scaling has ended

What’s next?

- Transistors
- Clock speed
- Power
- Perf / clock
Many-core tiled architectures

Motivation: increased performance without clock rate increase or smaller devices.

Requires changes in programming paradigm.

2010s: multi-core, GPGPU
Special purpose computation: Anton

(Shaw et al., 2014)
Can be implemented using:

- Silicon
- Software
- Neural network architectures!

<table>
<thead>
<tr>
<th>Motif</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Dense linear algebra</td>
<td>Matrix multiply (GEMM)</td>
</tr>
<tr>
<td>2 Sparse linear algebra</td>
<td>SpMV</td>
</tr>
<tr>
<td>3 Spectral methods</td>
<td>FFT</td>
</tr>
<tr>
<td>4 N-Body methods</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>5 Structured grids</td>
<td>Lattice Boltzmann</td>
</tr>
<tr>
<td>6 Unstructured grids</td>
<td>CFD</td>
</tr>
<tr>
<td>7 Map-Reduce</td>
<td>Expectation</td>
</tr>
<tr>
<td>8 Combinational logic</td>
<td>Encryption, hashing</td>
</tr>
<tr>
<td>9 Graph traversal</td>
<td>Decision trees, quicksort</td>
</tr>
<tr>
<td>10 Dynamic programming</td>
<td>Forward-backward</td>
</tr>
<tr>
<td>11 Bactrack, branch and bound</td>
<td>Constraint satisfaction</td>
</tr>
<tr>
<td>12 Graphical models</td>
<td>HMM, Bayesian</td>
</tr>
<tr>
<td>13 Finite state machines</td>
<td>Compilers</td>
</tr>
</tbody>
</table>

(Asanovic et al., 2006)
Summary

• Computers are tools for solving problems of their time
  • Was: Coding, calculation, graphics, web
  • Today: Learning and Inference on data
  • Deep learning as a computational paradigm
    • Custom architecture can do vastly better
  • We are hiring! Summer interns and full time.