Computation
Machine Learning Hardware

- CPUs
- GPUs
- FPGAs
- Other accelerators
Machine Learning Project

Training Data → Features → Training → Learned model

Test Data → Features → Prediction

Training Labels
Machine Learning hardware

- Speed up each block of the pipeline for example, matrix-matrix multiplication, convolution

- Data or memory paths for machine learning work example: caching

- Application-specific functional units
Processing

- **CPU** is good at executing few complex operations.

- *In ML* most of the processing involves matrix multiplication.

- **Lots of small calculations.**

- **GPU** is well suited for those kind of computations.
Processing

- GPU utilizes parallel architecture.

- It is very good at handling many sets of very simple instructions.
GPU’s
CPU vs. GPU

18 cores

2560 cores
## CPU vs. GPU

<table>
<thead>
<tr>
<th></th>
<th>CPU i9 Xseries</th>
<th>GeForce GTX 1080</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>18 (36 threads)</td>
<td>2560</td>
</tr>
<tr>
<td><strong>Clock Speed (GHz)</strong></td>
<td>4.4</td>
<td>1.6G</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>Shared</td>
<td>8GB</td>
</tr>
<tr>
<td><strong>Price ($)</strong></td>
<td>1799</td>
<td>549</td>
</tr>
</tbody>
</table>
GPU programming

• CUDA
  • C-like code that runs on GPU
  • Other APIs: cuBLAS, cuFFT, cuDNN, etc

• OpenCL
  • Similar to CUDA, but runs on CPU’s as well
  • usually slower
Frameworks

- Caffe (Berkeley)
- Torch (NYU / Facebook)
- Theano (University of Montreal)
Frameworks

- Caffe (Berkeley)
- Caffe2 (Facebook)
- Torch (NYU / Facebook)
- PyTorch (Facebook)
- Theano (University of Montreal)
- TensorFlow (Google)
- Paddle (Baidu)
- CNTK (Microsoft)
- MXNet (Amazon)
Deep Learning Frameworks

- Computational graphs
- Gradient computation
- Run on GPU seamlessly