ECE 598NS: Machine Learning in Silicon
Fall 2017

Introduction

Naresh Shanbhag
Professor
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
• motivation – why machine learning in silicon?
• what is inference? elements of an inference engine
• communication-inspired design of machine learning in silicon
• machine learning on stochastic nanoscale fabrics
• ECE 598NS syllabus
Is a Smart Phone Intelligent?

• what is intelligence?

Full Definition of intelligence

1. a (1) : the ability to learn or understand or to deal with new or trying situations; also: the skilled use of reason (2) : the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (as tests)

b Christian Science: the basic eternal quality of divine Mind
c: mental acuteness: shrewdness

2. a: an intelligent entity; especially: angel
   b: intelligent minds or mind <cosmic intelligence>

3. : the act of understanding: comprehension

4. a: information, news
   b: information concerning an enemy or possible enemy or an area; also: an agency engaged in obtaining such information

5. : the ability to perform computer functions
intelligence – a working definition:

intelligent systems are those that learn to make accurate decisions using limited data and computational resources, and are able to adapt to changes in environment.

requires extracting information from data
Applications Challenge
To extract relevant Information efficiently from Data

recognition, prediction, classification
decision-making, learning, adaptation
Intelligence in Computing Platforms

The Cloud
“Intelligence in the Cloud”

The Swarm
“Intelligence at the Edge”

Autonomous Platforms
(robots, UAVs, driverless cars)

Human-centric Platforms
(Wearables, biomedical sport/fitness)

Intelligence in Computing Platforms

resource-constrained platforms
A game-changing result

AlphaGo’s masters taught it the game, but an electrifying match shows what the computer may have to teach humans

Mar 19th 2016 | From the print edition

IT WAS not quite a whitewash, but it was close. When DeepMind, a London-based artificial intelligence (AI) company bought by Google for $400m in 2014, challenged Lee Sedol to a five-game Go match, Mr Lee—one of the best human players of that ancient and notoriously taxing board game—confidently predicted that he would win 5-0, or maybe 4-1.

He was right about the score, but wrong about the winner. The match, played in Seoul to
The Energy Cost of Intelligence

ICML 2016 Workshop on *On-device Intelligence*

Energy-efficient Machine Learning in Silicon: A Communications-inspired Approach

Narvsh Shanbhag  
University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

- information extraction costs much energy!
- Google DeepMind – 1202 CPUs and 176 GPUs
- roughly $10,000 \times$ more energy than the $\sim 20W$ human brain
- information extraction costs much energy!
Market Buzz

MICROSOFT EXTENDS FPGA REACH FROM BING TO DEEP LEARNING

August 27, 2015  Timothy Prickett Morgan

After three years of research into how it might accelerate its Bing search engine using field-programmable gate arrays (FPGAs), Microsoft came up with a scheme that would let it lash Stratix V devices from Altera to the two-socket server nodes in the minimalist Open Cloud Servers that it has designed expressly for its hyperscale datacenters. These CPU-FPGA hybrids were rolled out into production earlier this year to accelerate Bing page rank functions, and Microsoft started hunting around for other workloads with which to juice with...

QUALCOMM

Introducing Qualcomm® Zeroth® Platform

Qualcomm Technologies’ first cognitive computing platform designed for on-device intelligence.

Learning and adapting to the needs of the user.

Visual perception • Intelligent connectivity • Intuitive security • Always-on awareness
Immersive multimedia • Speech and audio recognition • Natural interaction

Find out more about Zeroth Platform at qualcomm.com/zeroth

While other companies are arguing about whether GPUs, FPGAs, or VPUs are better suited for machine learning, Google came out with the news that it has been using its own custom-built Tensor Processing Unit (TPU) for over a year, achieving a claimed 10x...
Maximize intelligence-per-unit-volume

- Low-latency cognitive decision-making
- Continuous on-device learning

The New Game
On-Device Intelligence

Under stringent resource constraints

- on energy, storage, computational capacity

ICML 2016 Workshop on *On-device Intelligence*
Dominant Source of Energy Cost – System Level

massive data movement

data movement between platforms

DATA

TX/RX

sensor

TX/RX

processor

memory

INFORMATION

The Cloud

data movement within platforms
Dominant Source of Energy Cost – Device Level

use of deterministic device fabrics

- low defect density, delay of CMOS and spin gates, RRAM resistance, and other physical parameters are random variables
- high energy costs of suppressing nanoscale randomness
Energy-efficient Machine Learning in Silicon

• today’s approach → treat machine learning as another problem in computing
• our ECE 598NS approach → communications-inspired
  ─ Treat a communication receiver as an inference kernel.
• repurpose the **communications IC design techniques** for machine learning on deterministic nanofabrics
• study Shannon-inspired **statistical error compensation** techniques for machine learning on stochastic nanofabrics
• stochastic nanofabrics: circuits that operate at the limits of energy efficiency and robustness
• motivation – why machine learning in silicon?
• what is inference? elements of an inference engine
• communication-inspired design of machine learning in silicon
• machine learning on stochastic nanoscale fabrics
• ECE 598NS syllabus
Attributes of an Inference System

Data Representation
(features)
- objects, vectors, signals, RVs

Knowledge Representation
- hyperplanes, clusters, ...

Inference Task
- classification
- regression
- clustering
- density estimation

Learning Method

Architecture

Application
- Task 1
- Task 2
- vision, EEG analysis, navigation, surveillance
A Classifier Architecture

- a system for supervised learning for classification task
- key components:
  - data set: \((X, Y)\)
  - feature extractor (FE): preprocesses \(X\) to improve accuracy & speed-up classification
  - classifier makes decisions
  - trainer: learns parameters of FE and classifier to maximize accuracy

misclassification rate (accuracy):
\[
Pr\{\hat{Y} \neq Y\}
\]
# Building Blocks of Inference Engines

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM/Adaline</td>
<td>DP, slicer</td>
</tr>
<tr>
<td>Poly. kernel SVM</td>
<td>DPE, slicer</td>
</tr>
<tr>
<td>RBF kernel SVM</td>
<td>sum-of-square (also DP), exponential, slicer</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>DP, logistic function ( f(x) = \frac{1}{1 + e^{-x}} )</td>
</tr>
<tr>
<td>CNN</td>
<td>DPE, ReLU function ( f(x) = \max(0, x) )</td>
</tr>
<tr>
<td>Random forest</td>
<td>comparator array, look-up table (LUT), voter</td>
</tr>
<tr>
<td>KNN/K-means</td>
<td>sum-of-square (also DP), add-compare-select</td>
</tr>
<tr>
<td>PCA</td>
<td>DPE</td>
</tr>
<tr>
<td>Other hand-made feature extractor</td>
<td>Filter bank, FFT, L1/L2 norm, SAD, cross correlation, random projection (also DP)</td>
</tr>
</tbody>
</table>

- Common feature extractor: filter, DP, DPE, FFT, various distance calculation block (L1/L2/SAD), exponential/log, comparator array
- Common classifier: slicer, voter, LUT, add-compare-select, logistic unit, ReLU


**Supervised learning**
- AODE
- Artificial neural network
  - Backpropagation
  - Autoencoders
  - Hopfield networks
  - Boltzmann machines
  - Restricted Boltzmann Machines
  - Spiking neural networks
- Bayesian statistics
  - Naive Bayes classifier
  - Bayesian network
  - Bayesian knowledge base
- Case-based reasoning
- Decision trees
- Inductive logic programming
- Gaussian process regression
- Gene expression programming
- Group method of data handling (GMDH)
- Learning Automata
- Learning Vector Quantization
- Logistic Model Tree
- Minimum message length (decision trees, decision graphs, etc.)
- Lazy learning
- Instance-based learning
  - Nearest Neighbor Algorithm
  - Probably approximately correct learning (PAC) learning
  - Ripple down rules, a knowledge acquisition methodology
  - Symbolic machine learning algorithms
- Subsymbolic machine learning algorithms
- Support vector machines
- Random Forests
- Ensembles of classifiers
  - Bootstrap aggregating (bagging)
  - Boosting (meta-algorithm)
- Ordinal classification
- Regression analysis
- Information fuzzy networks (IFN)
- Conditional Random Field
- **Stochastic Gradient Descent:**
  - Adaline, LMS

**Single-Stage Classifiers:**
- Support vector machines (SVM), non-linear SVM, Logistic Regression, Perceptron

**Deep Learning:**
- Deep belief networks
- Restricted Boltzmann Machines
- Convolutional Neural Networks

**Unsupervised learning**
- Artificial neural network
- Data clustering
- Expectation-maximization algorithm
- Self-organizing map
- Radial basis function network
- Vector Quantization
- Generative topographic map
- Information bottleneck method
- IBSEAD
- **Association rule learning**
  - Apriori algorithm
  - Eclat algorithm
  - FP-growth algorithm
- **Hierarchical clustering**
  - Single-linkage clustering
  - Conceptual clustering
- **Partitional clustering**
  - K-means algorithm
  - Fuzzy clustering
  - DBSCAN

**Reinforcement learning**
- Temporal difference learning
- Q-learning
- Learning Automata
- Monte Carlo Method
- SARSA

**Deep learning**
- Deep belief networks
- Deep Boltzmann machines
- Deep Convolutional neural networks
- Deep Recurrent neural networks

---

**ECE 598NS**

**supervised**

**Stochastic Gradient Descent:**
Adaline, LMS

**Single-Stage Classifiers:**
Support vector machines (SVM), non-linear SVM, Logistic Regression, Perceptron

**Deep Learning:**
Deep belief networks
Restricted Boltzmann Machines
Convolutional Neural Networks

**Ensembles of Classifiers:**
ADABoost, Random Forest

---

**ECE ILLINOIS**
Department of Electrical and Computer Engineering
A Micro-Power EEG Acquisition SoC With Integrated Feature Extraction Processor for a Chronic Seizure Detection System

Naveen Verma, Member, IEEE, Ali Shoeb, Jose Bohorquez, Member, IEEE, Joel Dawson, Member, IEEE, John Guttag, and Anantha P. Chandrakasan, Fellow, IEEE

- seizure detection from EEG waveforms for epilepsy
- trained inference engine

Fig. 1. 18-channel EEG showing onset of patient seizure (ictal); electrical onset occurs 7.5 sec before the clinical onset, which is characterized by muscle reflexes causing the large excursion artifacts.
• Local feature extraction to achieve energy efficiency. Need to transmit features (low-volume) vs. data. \((1.8mW \rightarrow 0.12mW)\)
A Complete Inference System

- method to acquire and store raw data (IA & ADC)
- represent data in a form amenable to inference (FE proc.)
- inference task to be implemented (SVM)
- method to learn the inference model parameters (Trainer)
- efficient architecture and platform realization (IC)
Feature Extraction Engine

- information in an EEG signal resides in the 0-20Hz band.
- FE: 7 channel filter bank, 3Hz passband, 1.5Hz stop band, filter order of 44 order, BW of 18+1.5=19.5Hz
- input sampling rate: 600S/s with 12b ADC, input data-rate of $12 \times 600 = 7200$ bits/s.
- output feature = 112-bit (7 filters $\times$ 16-bit filter output) vector (2 second observation window)
- feature rate 56 bits/s $\rightarrow$ total data volume reduction of $129X$
• motivation – why machine learning in silicon?
• what is inference? elements of an inference engine
• communication-inspired design of machine learning in silicon
• machine learning on stochastic nanoscale fabrics
• ECE 598NS syllabus
• Pop quiz – What is the world’s most popular learning algorithm deployed “on-device” today?
• Answer: the least mean-square (LMS) algorithm (Widrow-Hoff, 1960).
  – Originally used to train Adaptive Linear Neuron (ADALINE) & MADALINE
• Used in communication receivers since the mid-’60’s
  – Channel estimation, echo/crosstalk cancellation, equalization (Got a cell-phone, have LMS!)
• LMS = SGD applied to a linear combiner to minimize MSE
• SGD is the workhorse of deep learning networks today
• design principles for energy-efficient, high-throughput, capacity achieving, communication ICs with learning capabilities are well-established → how to leverage these?

Ph.D. thesis ‘93

On-chip Learners
(adaptive equalizers)

51.84 Mb/s
Very high-speed DSL receivers
(AT&T)

12.5 Gb/s
Long-haul optical receivers
(Intersymbol Comm.)

Algorithms
Architectures
Integrated Circuits

Viterbi equalizer with
On-chip Learner
(channel estimator)
Classifiers in Communication Links

• Classifiers used in communications links are referred to as detectors.
• An forward error-control (FEC) decoder is a detector that bases its decision on a block of channel symbols $\hat{c}[k]$
• A slicer is a detector that bases its decisions on a single observation $y_R[k]$. 
• **transmitter imposed structure** on models of the input data (signal) and noise, e.g. additive noise model:

\[ y_R[k] = c[k] + n[k] \]

• imposed structure on data makes the **classifier computationally efficient and accurate** → key distinction between machine learning and communications
A communication receiver is an inference engine.

- **Transmitter**
  - Encoded information $Y$
  - Modulated signal $X$

- **Channel**
  - Noise $N$
  - Received signal $X$

- **Receiver**
  - Demodulated signal $\hat{X}$
  - Recovered information $\hat{Y}$

- **Channel Estimator**
  - Training sequence $Y$

- **Heavily engineered statistics**
  - Time series, “unlimited”

- Infers information $Y$ from noisy received data $X$
- Needs to be trained on-line/chip to do this – and we know how → repurpose these for on-device intelligence.
• **Strong connection** between algorithms, arch., circuits

• **Reduce algorithmic complexity & precision** subject to accuracy requirements – has been done analytically

\[
B_F \geq \frac{1}{2} \log_2 \left( \frac{N \sigma_x^2}{12 \alpha \sigma_y^2} \right) + \frac{SNR_{fl}(dB)}{6}
\]

\[
B_{WUD} \geq \frac{1}{2} \log_2 \left( \frac{1}{\mu^2 \sigma_y^2 \sigma_x^2} \right) + \frac{SNR_{fl}(dB)}{6}
\]


• **Use systematic algorithm-to-architecture mapping techniques**, e.g., systolic transformations → machine learning DFGs tend to be regular

• motivation – why machine learning in silicon?
• what is inference? elements of an inference engine
• communication-inspired design of machine learning in silicon
• machine learning on stochastic nanoscale fabrics
• ECE 598NS syllabus
• Repurposing will get us a good baseline design (and avoid reinvention) but .... the big question is:

How do we design intelligent platforms operating at the limits of energy efficiency, throughput & information density.........?
On-device intelligence

- Cognitive, decision-making
- Continuous on-device learning

How do we operate at the Fundamental Limits?

Max. intelligence-per-unit-volume
St. resource constraints

- energy, storage, computational capacity

Resource constraints

Claude Shannon

1948
Obtained Fundamental Limits for Communication Systems

50yrs later
Communication systems operate at the Limits with Capacity Achieving Codes
Limiting behavior of nanofabrics is **stochastic**

- CMOS in near-threshold voltage (NTV) is unreliable (also very slow)
- Spin devices need to operate in the stochastic regime in order to be competitive with CMOS
  - Stochastic regime → trade-off between error rate, delay and energy
- Many others...

---

[Diagram showing CMOS vs. RRAM performance, with stochastic and deterministic regions highlighted.]

[Wong]
Lower Bounds on Power-Dissipation for DSP Algorithms

Naresh R. Shanbhag
ISLPED 1996 Monterey CA USA
0-7803-3571-8/96/$5.00 ©1996

Energy-Efficient Signal Processing via Algorithmic Noise-Tolerance *

Rajamohana Hegde and Naresh R. Shanbhag
ISLPED99, San Diego, CA, USA
©1999 ACM 1-58113-133-X/99/0008..$5.00

Reliable Low-Power Design in the Presence of Deep Submicron Noise

Naresh Shanbhag K. Soumyanath Samuel Martin
ISLPED '00, Rapallo, Italy.
Copyright 2000 ACM 1-58113-190-9/00/0007...$5.00.
Shannon-inspired Statistical Error Compensation (SEC)

- Treat computation on stochastic fabric as a noisy channel
- Leverage statistical estimation, detection, and inference techniques
  - explicit or implicit use of signal and error statistics

[Hedge & Shanbhag, IEEE Transactions on VLSI’01, IEEE Journal of Solid-State Circuits’04]
Statistical Error Compensation Techniques

**Algorithmic noise-tolerance (ANT)**

\[
y_1 = y_o + \eta \\
y_2 = y_o + \varepsilon
\]

[ISLPED99, CICC01, JSSC04, TVLSI04, TVLSI08, JSSC13]

**Stochastic sensor NOC (SSNOC)**

[TVLSI10, CICC11, TVLSI14]

**Soft NMR**

[Trans. Computers’12]

**Likelihood Processing**

[Trans. on Multimedia’13]
<table>
<thead>
<tr>
<th>System Description</th>
<th>Energy Reduction</th>
<th>Error Rates</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>256-tap PN code detection filter in 180nm CMOS</td>
<td>5.8X</td>
<td>$P_{\text{det}} &gt; 90%$ with error rates $&lt; 86%$</td>
<td>[Kim, Shanbhag, et al., CICC 2012]</td>
</tr>
<tr>
<td>Subthreshold ECG classifier in 45nm CMOS</td>
<td>28% (wrt MEOP)</td>
<td>$P_{\text{det}} &gt; 95%$ with error rates $&lt; 58%$</td>
<td>[Abdallah, Shanbhag, IEEE Journal of Solid-State Circuits, 2013]</td>
</tr>
</tbody>
</table>

- Ability to handle high computational error rates demonstrated in prototype inference ICs → apply to deep learners
A Systems-driven approach to extend Moore’s Law into the deep Nanoscale regime

by developing Shannon & Brain-inspired statistical information processing principles, architectures, prototypes

Director: Naresh Shanbhag
Associate Director: Andrew Singer

Illinois (LEAD), Berkeley, Stanford, UCSD, UCSB, Michigan, CMU, Princeton, Cornell, MIT

[2013–’17]
The New Game:
depth learning on mobile platforms

Maximize intelligence-per-unit-volume
- Cognitive, decision-making
- Continuous on-device learning

Under stringent resource constraints
- on energy, storage, computational capacity

IT WAS not quite a whitewash, but it was close. When DeepMind, a London-based artificial intelligence (AI) company bought by Google for $400m in 2014, challenged Lee Sedol to a five-game Go match, Mr Lee—one of the best human players of that ancient and notoriously taxing board game—confidently predicted that he would win 5-0, or maybe 4-1.

He was right about the score, but wrong about the winner. The match, played in Seoul to
A Nanotechnology-Inspired Grand Challenge for Future Computing

OCTOBER 20, 2015 AT 6:00 AM ET BY LLOYD WHITMAN, RANDY BRYANT, AND TOM KALIL

Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.

https://www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing
• motivation – why machine learning in silicon?
• what is inference? elements of an inference engine
• communication-inspired design of machine learning in silicon
• machine learning on stochastic nanoscale fabrics
• ECE 598NS syllabus
Course Announcement
ECE 598 - Machine Learning in Silicon (Fall 2017)

Instructor: Naresh R. Shanbhag, http://shanbhag.ece.illinois.edu
Credit: 4 hours

Prerequisites: ECE 482 or equivalent. Students should be familiar with programming in MATLAB. HDL (VHDL/Verilog) programming experience is desirable.

Textbook: Instructor notes and assigned technical papers.

Time and Place: 11:00 -12.20, MW, 2013 ECEB
• Grading:
  – Three homework assignments (25%)
  – One paper presentation (25%)
  – Course project (50%)

Instructor Office Hours:

Wednesdays: 2pm-3pm, 414 CSL

Contact Prof. Shanbhag at shanbhag@illinois.edu, if consultation at a different time is needed.

Course Web-Page: http://courses.ece.uiuc.edu/ece598ns/fa2017
TAs: Ameya Patil (adpatil2@illinois.edu) and Charbel Sakr (sakr2@illinois.edu);
TA Office Hours: Patil, Sakr (2-3PM on Fridays);
Summary

• we live in a data-rich world → need inference kernels (extract information from data) → energy problem
• Moore’s Law has slowed down → robustness (cost) problem
• a communication receiver is an inference kernel
• communication-inspired approach to the design of machine learning in silicon – approach advocated in ECE 598NS
  – machine learning on deterministic fabrics
  – machine learning on stochastic fabrics (Shannon-inspired)
https://courses.engr.illinois.edu/ece598ns/fa2017

http://shanbhag.ece.uiuc.edu