ECE 8893 B
Hardware Acceleration for Machine Learning
Spring 2019

INTRODUCTION

Tushar Krishna
Assistant Professor
School of Electrical and Computer Engineering
Georgia Institute of Technology

tushar@ece.gatech.edu
INTRODUCTIONS!

Background
- PhD from MIT in EECS (2013)
- Researcher at Intel (2014-15)
- Georgia Tech (2015 - present)

Office: Klaus 2318
Office Hours:
Tues: 1:30 – 3PM in Klaus 2318

Research Interests
- Computer Architecture
- Interconnection Networks
- Network-on-Chip
- Deep Learning Accelerators
- Continuous Learning Systems
AGENDA

- Course Motivation
- Course Logistics
- Introduction to Deep Learning
COURSE MOTIVATION
THE DREAM!
“AI is the new electricity” – Andrew Ng
WHAT IS THIS COURSE ABOUT?

- How to leverage the rise of ML\(^1\) to fuel the end of ML\(^2\)
  - 1: Machine Learning
  - 2: Moore’s Law

Source: Jeff Dean, Google
TAKEAWAYS FROM THIS COURSE

- Separating the Hype from the Opportunity
  - Why is everyone and their uncle working on Machine Learning?

- Computer Architecture perspective
  - The role of HW Acceleration in the AI/ML Boom
    - Understand Compute and Memory behavior of Deep Learning Workloads
    - Understand Limitations of Current Platforms (CPU and GPU)
    - What about Moore’s Law?
  - Opportunities and Challenges with custom HW Accelerators
    - Techniques to reduce computation
    - Techniques to reduce data movement
    - Techniques to reduce memory overhead
    - Performance vs Energy vs $$

- ML + Systems Perspective
  - Co-Design of DNN and Mapping/Dataflow and Hardware
  - Emerging Algorithms and Emerging Technologies
  - Neuromorphic (Brain-like) Computing
**Course Information**

**Course Website**
- [http://tusharkrishna.ece.gatech.edu/teaching/hml_s19/](http://tusharkrishna.ece.gatech.edu/teaching/hml_s19/)
- All additional readings will be posted here!

**Canvas**
- **Lecture Slides will be posted here**
- **Lab Assignments will be posted and submitted here**
- **Project Reports will be submitted here**

**Piazza**
- Access via Canvas
- **Use for questions related to the lab assignments**
  - Try to answer each other’s questions
  - **Only part-time TA in the course**
  - Do not upload code!
  - If no one is able to answer within a day I will try and respond
- **Discussions on paper readings also encouraged**
TEACHING ASSISTANT (PART-TIME)

Anand Samajdar
anandsamajdar@gatech.edu

Office Hours: TBA
<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Due [Friday]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Jan 7 - )</td>
<td></td>
<td></td>
<td>Lab 1 – Part A</td>
</tr>
<tr>
<td>2</td>
<td>(Jan 14 - )</td>
<td></td>
<td></td>
<td>Lab 1 – Part B</td>
</tr>
<tr>
<td>3</td>
<td>(Jan 21 - )</td>
<td>MLK Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(Jan 28 - )</td>
<td></td>
<td></td>
<td>Lab 2</td>
</tr>
<tr>
<td>5</td>
<td>(Feb 4 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(Feb 11 - )</td>
<td></td>
<td></td>
<td>Lab 3</td>
</tr>
<tr>
<td>7</td>
<td>(Feb 18 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(Feb 25 - )</td>
<td></td>
<td></td>
<td>Lab 4</td>
</tr>
<tr>
<td>9</td>
<td>(Mar 4 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>(Mar 11 - )</td>
<td></td>
<td>Proposal Ppt</td>
<td>Lab 5</td>
</tr>
<tr>
<td>11</td>
<td>(Mar 18 - )</td>
<td>Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>(Mar 25 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>(Apr 1 - )</td>
<td></td>
<td></td>
<td>Milestone 1</td>
</tr>
<tr>
<td>14</td>
<td>(Apr 8 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>(Apr 15 - )</td>
<td></td>
<td></td>
<td>Milestone 2</td>
</tr>
<tr>
<td>16</td>
<td>(Apr 22 - )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>(Apr 29 - )</td>
<td>Final Presentations</td>
<td></td>
<td>Final Report</td>
</tr>
</tbody>
</table>

**Notes:**
- GT Holiday
- Instructor Travel

**Calendar Notes:**
- MLK Day
- Spring Break
- Milestone 1
- Milestone 2
- Final Presentations
- Final Report
I tried offering the course as 2-3-3 (instead of 3-0-0) to emphasize lab assignments

Department gave us a 50 minute slot
- But I intend to use it as a 1hr 20min slot (3-0-0) course
- 6 to 7:20 PM
LABS (TENTATIVE)

- Lab 1 – Keras (w Tensor Flow backend)
  - Running MLP and CNN on CPU and GPU
  - Part A (CPU) due this Friday by 1 PM
    - Gives you time to drop course if you think you do not have the background or bandwidth
  - Part B (GPU) due next Friday at Midnight

- Lab 2 – Design-space Exploration of TPU-like Accelerator

- Lab 3 – Verilog Implementation of TPU-like Accelerator

- Lab 4 – Dataflow (Mapping) Exploration

- Lab 5 – Hardware Design-space Exploration for Spatial Accelerator
<table>
<thead>
<tr>
<th>Item</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labs</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Paper Critiques</strong></td>
<td>10% [Best of 10]</td>
</tr>
<tr>
<td>Project - Proposal</td>
<td>5%</td>
</tr>
<tr>
<td>Project - Milestones</td>
<td>10%</td>
</tr>
<tr>
<td>Project - Presentation</td>
<td>10%</td>
</tr>
<tr>
<td>Project - Final Report</td>
<td>15%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
</tr>
</tbody>
</table>
HOW TO CONTACT ME

- Piazza for clarifications on labs

- Email for Specific Questions
  - E.g., project proposal ideas

- Office Hours
  - After Class
  - Tuesday 1:30 – 3PM in Klaus 2318
  - Setup appointment via email
REQUIRED BACKGROUND

- **Systems**
  - Computer Architecture
  - Digital Hardware Design
    - Verilog
    - CAD Tools
    - FPGA (not necessary but a plus)

- **Machine Learning**
  - Basic understanding of Deep Neural Networks
    - CNN
    - GEMM
    - Training vs Inference
  - Understanding why it works (the Math behind ML) not required
    - Understanding how well it works (performance and energy-efficiency) will be the focus
WHAT IS EXPECTED OF YOU?

- **Self-drive and Motivation**
  - New course – one of its kind
    - No textbook!
    - I will be learning many of the topics with you
  - Many resources online
    - Too many resources online

- **Presentations and Discussions**
  - Class Participation
  - Project Proposal Presentation
  - Final Presentation
WHAT IS EXPECTED OF YOU?

- **Research Papers**
  - You will be reading, understanding and critiquing ~20 research papers
  - Topics: ML Algorithms / Architecture / Emerging Technologies
- **Critique (1 page)**
  - Short Summary + 2 strengths + 2 weaknesses + 1 suggested improvement
  - Due at the beginning of class
WHAT IS EXPECTED OF YOU?

- **Hands-on System Building and Hacking**
  - Labs due every 2 weeks
    - 2-3-3 course
  - Each Lab will require installing, learning and working with a new framework
    - Keras + Tensor Flow (DNN Model)
    - MAESTRO (Dataflow Mapper)
    - SCALE-Sim (Microarch Simulator)
    - Verilog + CAD Tools – ASIC and FPGA
  - **Project** – implement and evaluate a new idea within 4-6 weeks
ADDITIONAL RESOURCES

- **Deep Learning Textbooks and Courses**
  - [https://www.deeplearning.ai/](https://www.deeplearning.ai/) (Andrew Ng)

- **DNN Packages (examples, tutorials, pre-trained models)**
  - Keras
  - Caffe
  - Theano
  - TensorFlow
  - PyTorch
  - Matconvnet

- **Sister Courses on ML Hardware Acceleration**
  - MIT 6.S082/6.888: Hardware Architectures for Deep Learning (Joel Emer and Vivienne Sze)
  - Stanford CS 217: Hardware Accelerators for Machine Learning (Kunle Olukotun and Ardavan Pedram)
INTRO TO DEEP LEARNING

Acknowledgment:
Slides adapted from
“Hardware Architectures for Deep Neural Networks”, ISCA Tutorial 2017
Joel Emer and Vivienne Sze
Artificial Intelligence

“The science and engineering of creating intelligent machines”
- John McCarthy, 1956
“Field of study that gives computers the ability to learn without being explicitly programmed”

– Arthur Samuel, 1959
An algorithm that takes its basic functionality from our understanding of how the brain operates.
HOW DOES THE BRAIN WORK?

- The basic computational unit of the brain is a **neuron** → 86B neurons in the brain

- Neurons are connected with nearly $10^{14} - 10^{15}$ synapses

- Neurons receive input signal from **dendrites** and produce output signal along **axon**, which interact with the dendrites of other neurons via **synaptic weights**

- Synaptic weights – learnable & control influence strength
SPIKING-BASED MACHINE LEARNING

Artificial Intelligence

Machine Learning

Brain-Inspired

Spiking

Slide Courtesy: Joel Emer and Vivienne Sze
SPIKING ARCHITECTURE

- Brain-inspired - Integrate and fire
- Example: IBM TrueNorth

[Merolla et al., Science 2014; Esser et al., PNAS 2016]


Slide Courtesy: Joel Emer and Vivienne Sze
NEURON: WEIGHTED SUM

\[ x_0 \text{ synapse} \quad w_0 \]

axon from a neuron

dendrite

\[ w_0 x_0 \quad \sum_{i} w_i x_i + b \rightarrow f \left( \sum_{i} w_i x_i + b \right) \]

cell body

activation function

output axon

Image Source: Stanford
NEURAL NETWORK

[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze
NEURAL NETWORK: MULTIPLE WEIGHTED SUMS

Each synapse has a weight for neuron activation

\[ Y_j = \text{activation}\left( \sum_{i=1}^{3} W_{ij} \times X_i \right) \]

[Image Source: Stanford]
Weight Sharing: multiple synapses use the same weight value

\[ Y_j = \text{activation}\left(\sum_{i=1}^{3} W_{ij} \times X_i \right) \]
NEURAL NETWORK TERMINOLOGY

Layer 1

L1 Neuron inputs
e.g. image pixels

[Image Source: Stanford]

L1 Neuron outputs
a.k.a. Activations

input layer

hidden layer

output layer

Slide Courtesy: Joel Emer and Vivienne Sze
NEURAL NETWORK TERMINOLOGY

L2 Input Activations

Layer 2

L2 Output Activations

input layer

hidden layer

output layer

[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze
NEURAL NETWORK TERMINOLOGY

A **layer** can refer to a set of activations or a set of weights. In this class, we use **layer** to refer to a set of weights.

“2-layer Neural Net”, or “1-hidden-layer Neural Net”

“3-layer Neural Net”, or “2-hidden-layer Neural Net”

[Image Source: Stanford]

*Slide Courtesy: Joel Emer and Vivienne Sze*
**NEURAL NETWORK TERMINOLOGY**

**Fully-Connected**: all i/p neurons connected to all o/p neurons

**Sparsely-Connected**

[Image Source: Stanford]

[Image Courtesy: Joel Emer and Vivienne Sze]
NEURAL NETWORK TERMINOLOGY

Feed Forward

Feedback

input layer

hidden layer

output layer

[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze
DEEP LEARNING

Artificial Intelligence

Machine Learning

Brain-Inspired

Spiking

Neural Networks

Deep Learning

Slide Courtesy: Joel Emer and Vivienne Sze
POPPULAR TYPES OF DEEP NEURAL NETWORKS

- Fully-Connected NN
  - feed forward, a.k.a. multilayer perceptron (MLP)

- Convolutional NN (CNN)
  - feed forward, sparsely-connected w/ weight sharing

Slide Courtesy: Joel Emer and Vivienne Sze
POPOULAR TYPES OF DEEP NEURAL NETWORKS

- Recurrent NN (RNN)
  - feedback

- Long Short-Term Memory (LSTM)
  - feedback + storage

Feed Forward

Feedback

input layer

hidden layer

output layer

Slide Courtesy: Joel Emer and Vivienne Sze
Inference

Convolutional Layers (Feature Extraction)

Summarize features

Intermediate features

“Klaus Advanced Computing Building”
DEEP LEARNING LANDSCAPE

Training → Inference

Labelled Datasets → ML Practitioner → Error

"Klaus Advanced Computing Building"

DNN Model (Topology + Weights)
WHY IS DEEP LEARNING HOT NOW?

Big Data Availability

New ML Techniques

GPU Acceleration

Facebook
350M images uploaded per day

YouTube
300 hours of video uploaded every minute

Walmart
2.5 Petabytes of customer data hourly

Convolutional Neural Network

Convolutional Neural Network

Slide Courtesy: Joel Emer and Vivienne Sze
Image Classification Task:
1.2M training images • 1000 object categories

Object Detection Task:
456k training images • 200 object categories
**IMAGENET: IMAGE CLASSIFICATION TASK**

**Top 5 Classification Error (%)**

- **Large error rate reduction due to Deep CNN**

<table>
<thead>
<tr>
<th>Year</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>29.1</td>
</tr>
<tr>
<td>2011</td>
<td>28.9</td>
</tr>
<tr>
<td>2012</td>
<td>16.9</td>
</tr>
<tr>
<td>2013</td>
<td>13.9</td>
</tr>
<tr>
<td>2014</td>
<td>8.9</td>
</tr>
<tr>
<td>2015</td>
<td>3.9</td>
</tr>
<tr>
<td>Human</td>
<td>1.9</td>
</tr>
</tbody>
</table>

**Hand-crafted feature-based designs** vs **Deep CNN-based designs**

[Russakovsky et al., IJCV 2015]

*Slide Courtesy: Joel Emer and Vivienne Sze*
GPU USAGE FOR IMAGENET CHALLENGE

- Top 5 Error Rate
- # of entries using GPUs

Year 2010: 28% error rate, 0 entries using GPUs
Year 2011: 26% error rate, 0 entries using GPUs
Year 2012: 16% error rate, 4 entries using GPUs
Year 2013: 12% error rate, 60 entries using GPUs
Year 2014: 7% error rate, 110 entries using GPUs

Slide Courtesy: Joel Emer and Vivienne Sze
DEEP LEARNING ON IMAGES

- Image Classification
- Object Detection
- Object Localization
- Image Segmentation
- Action Recognition
- Image Generation
DEEP LEARNING FOR SPEECH

• Speech Recognition
• Natural Language Processing
• Speech Translation
• Audio Generation
DEEP LEARNING ON GAMES

Google DeepMind AlphaGo

Slide Courtesy: Joel Emer and Vivienne Sze
MEDICAL APPLICATIONS OF DEEP LEARNING

• Brain Cancer Detection

Image Source: [Jermyn et al., JB0 2016]

Slide Courtesy: Joel Emer and Vivienne Sze
DEEP LEARNING FOR SELF-DRIVING CARS

Slide Courtesy: Joel Emer and Vivienne Sze
Machine Learning requires orders of magnitude more computation than other parts.
MATURE APPLICATIONS

• **Image**
  - Classification: image to object class
  - Recognition: same as classification (except for faces)
  - Detection: assigning bounding boxes to objects
  - Segmentation: assigning object class to every pixel

• **Speech & Language**
  - Speech Recognition: audio to text
  - Translation
  - Natural Language Processing: text to meaning
  - Audio Generation: text to audio

• **Games**
# EMERGING APPLICATIONS

- **Medical** (Cancer Detection, Pre-Natal)

- **Finance** (Trading, Energy Forecasting, Risk)

- **Infrastructure** (Structure Safety and Traffic)

- Weather Forecasting and Event Detection

http://www.nextplatform.com/2016/09/14/next-wave-deep-learning-applications/
OPPORTUNITIES

$500B Market over 10 Years!

Image Source: Tractica
Opportunities

From EE Times – September 27, 2016

"Today the job of training machine learning models is limited by compute, if we had faster processors we’d run bigger models...in practice we train on a reasonable subset of data that can finish in a matter of months. We could use improvements of several orders of magnitude – 100x or greater."

- Greg Diamos, Senior Researcher, SVAIL, Baidu
COMPUTATION PLATFORMS

- CPU
  - Intel, ARM, AMD…

- GPU
  - NVIDIA, AMD…

- Fine Grained Reconfigurable (FPGA)
  - Microsoft BrainWave

- Coarse Grained Programmable/Reconfigurable
  - Wave Computing, Plasticine, Graphcore…

- Application Specific
  - Neuflow, *DianNao, Eyeriss, Cnvlutin, SCNN, TPU, …
COMPUTATION PLATFORMS

Training ➔ Inference

HPC cluster
(CPU+GPU+FPGA)

Accelerators
(ASIC)

ARM Trillum
NVDLA
Apple Neural Engine
CambriconX
ShiDianNao
Eyeriss
CPU COMPUTE MODEL

Program → Compiler → Binary Program → Processor → Input Data

Compiler

Architecture

µArchitecture

Behavioral Statistics

Processed Data

© Tushar Krishna, School of ECE, Georgia Tech

Slide Courtesy: Joel Emer and Vivienne Sze
DNN Compute Model

Model (Shape) -> Mapper -> Configuration

Dataflow -> Implementation

DNN Accelerator

Input Activations -> Behavioral Statistics

Output Activations
CHALLENGES IN DESIGN AND DEPLOYMENT

DNN Model/Shape → Mapping (Dataflow) → Accelerator Microarchitecture → Energy → Runtime
PUBLICATIONS AT ARCHITECTURE CONFERENCES

- MICRO, ISCA, HPCA, ASPLOS

# of Publications over the Years

- Micro
- ISCA
- HPCA
- ASPLOS
DNN TIMELINE

- 1940s: Neural networks were proposed
- 1960s: Deep neural networks were proposed
- 1989: Neural network for recognizing digits (LeNet)
- 1990s: Hardware for shallow neural nets
  - Example: Intel ETANN (1992)
- 2011: Breakthrough DNN-based speech recognition
  - Microsoft real-time speech translation
- 2012: DNNs for vision supplanting traditional ML
  - AlexNet for image classification
- 2014+: Rise of DNN accelerator research
  - Examples: Neuflow, DianNao, etc.
NEXT TOPICS

- Basics of Deep Neural Networks
  - Feedforward - Multi-layer Perceptron
  - Feedback - Backpropogation

- Key Compute blocks in modern DNNs
  - GEMM
  - Convolutions

- Why DNN Accelerators?
  - Challenges with CPUs and GPUs
  - Deep Dive: Google TPU