Artificial Intelligence Hardware

Keras Intro

April 3, 2019

Huntsville AI - Facebook Group

What we cover:

- Application - how to solve a problem with a technology
- Theory - for those times when knowing “How” something works is necessary
- Social / Ethics - Human / AI interaction
- Brainstorming - new uses for existing solutions
- Hands on Code - for those times when you just have to run something for yourself
- Coworking Night - maybe have combined sessions with other groups to discuss application in their focus areas

https://www.facebook.com/groups/390465874745286/

About Ben...

Software Engineer (embedded) at Raytheon

Pursuing Masters in CS focusing on AI/ML

AI/ML Software Engineer at MTSI in 2 weeks.

Here’s Hardware stuff

AI/ML Hardware Concerns

(mostly deep learning)

Bottlenecks

- What’s holding your system back?

GPU

- Whys?
  - Most important!
    - … except when it’s not
  - Parallel computation
    - Especially artificial neural networks
  - Throughput focus
- Whats?
  - Memory
    - Amount
    - Speed
    - ECC (no)
  - Compute
    - FLOPS
      - FP64
      - FP32
      - FP16
  - Interconnects
  - Architecture

CPU

- Whys?
  - Sequential computation
  - Latency focus
- Whats?
  - Cores/Threads
    - Amount
    - Layout (TR/Epyc/Stacking)
  - Clock speed
  - PCIe
    - Lanes (bandwidth)
    - Revision
    - Latency
  - Architecture
    - Available instructions
  - Cache

RAM

- Whys?
  - Fast storage
  - Most running stuff lives here
- Whats?
  - Amount
  - Clock speed
    - More RAM -> Slower clocks (sort of)
  - Timings (tricky)

Network Interface

- Whys?
  - Big models
  - Big data
- Whats?
  - Bandwidth
  - Latency
  - RDMA

Storage

- Whys?
  - Big models
  - Big data
- Whats?
  - Bandwidth
  - Latency
  - Capacity

Keras Introduction

- What?
  - High level API (wrapper) for popular machine learning libraries.
- Why?
  - Less boilerplate code to get started. Fast prototyping.
- Officially high level API of tensorflow

Keras References

Created by Francois Chollet - @fchollet on Twitter

Lecture at Stanford on Keras

https://web.stanford.edu/class/cs20si/lectures/march9guestlecture.pdf

More intro material from TowardsDataScience:

https://towardsdatascience.com/introduction-to-deep-learning-with-keras-17c09e4f0eb2

Keras References

- The Sequential Model
  - Dead simple
  - Only for single-input, single-output, sequential layer stacks
  - Good for 70+% of use cases
- The functional API
  - Like playing with Lego bricks
  - Multi-input, multi-output, arbitrary static graph topologies
  - Good for 95% of use cases
- Model subclassing
  - Maximum flexibility
  - Larger potential error surface