Panel: Arithmetic Requirements for AI and Deep Learning

Chair: Eric Schwarz

Panelists

• Valentina Popescu	Intel - Nervana
• Martin Langhammer	Intel - FPGA Altera
• Stuart Oberman	NVidia
• Alberto Nannarelli	Technical University of Denmark

Applications

- What is the killer app of deep learning
  - Image recognition
  - Speech recognition
  - Watson like responses – conversational replies and arguments
  - Market intelligence
  - Designing arithmetic circuits?

- Where do we do training vs inference?
  - Is training batch processing and old?
  - Is inference best done in PDA – phone?

- Where do we do training?
  - CLOUD?

Computation

- What is the computational model – neural networks
  - Google TPU paper
    - 61% Multi-Layer Perceptrons (MLP)
    - 29% Long Short-Term Memory (LSTM) form of Recurrent Neural Network (RNN)
    - 5% Convolution Neural Networks (CNN)

- Also found 256 x 256 x 8 bit optimal

Operations

- Matrix Multiply for Neural Networks

Number formats for Training

- Traditionally use FP32
- Should we use FP16?
- Which format of FP16
  - IEEE 754 1/5/10
  - Google bfloat 1/8/7
- Should we use block or scaled fixed point?
  - Flex16+5

Other formats

- Posits (John Gustafson - Unums):
  - Sign, Regime, Exponent, Fraction
  - Regime (Re – jeam) uses a thermometer code
    - 1/65536, 1/256, 1/16, ¼, ½, 1, 2, 4, 16, 256, 655536
  - (-1)^S * R * 2^E * 1.F

- Stochastic Rounding (Suyog Gupta-2015)?

Standardization

- Should we standardize number format?
- Do we share data between systems?
- Does each company want to proprietary

Why Cloud?

- Google only making TPU available on their Cloud
- Microsoft FPGA Brainwave at BUILD conference for their Azure cloud
  - Also acquires Bonsai
- AWS working on own custom chips
- Facebook working on own custom chips

Other Applications

- Autonomous or self-driving Cars
- Tim Cook characterized the challenge of building autonomous vehicles as “the mother of all” AI projects.
- Would you explore a little bit on the point of view of autonomous driving? It needs huge amounts of computing power to interpret data from sensors and take action based on changing road conditions and traffic scenarios.
- Which arithmetic operations are most used in this type of application? Trigonometric operations are necessary to detect vehicle position and distance to objects.
    - Do the panelists have inputs on their experience working with these operations?
    - Which accuracy is required?
    - Did they contribute or could share their thoughts about standards that are in place today that define the properties of arithmetic operations for autonomous cars?

Implementation Bottlenecks

- Memory Bandwidth
- Clock rate
- Parallelism of 256 x 256 array