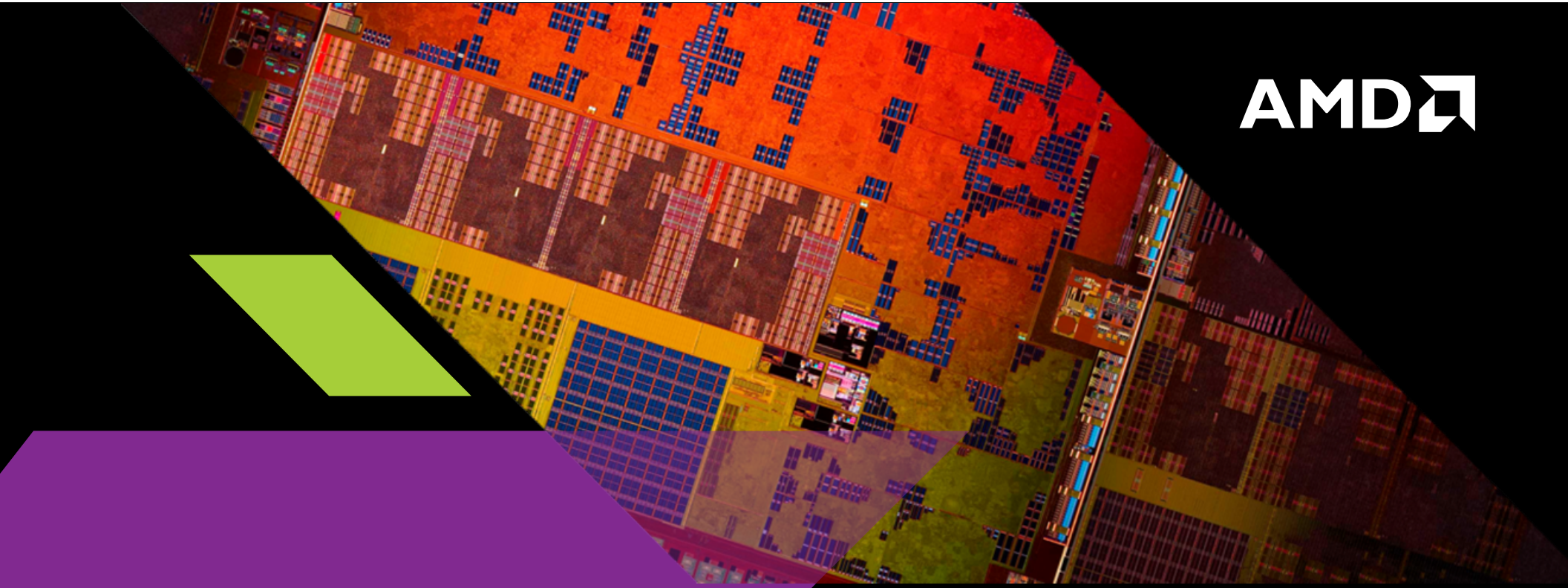


EXPLORATION OF NUMERICAL PRECISION IN DEEP NEURAL NETWORKS

NICHOLAS MALAYA AND ALLEN RUSH

6/19/17

ADVANCED MICRO DEVICES

SNAPSHOT OF THE SEMICONDUCTOR COMPANY

- CPUs [Ryzen]
  - One of two x86 suppliers
- GPUs [Radeon]
- Gaming Consoles:
  - Xbox One / Project Scorpio
  - PS4
  - WiiU
- APUs, Servers [EPYC], Supercomputers, etc.

Blue Waters Supercomputer

You have probably used our products

WHY IS AMD INTERESTED IN MACHINE LEARNING?

GPUS “KILLER APPLICATION”

- Simple Answer: interested in all computation

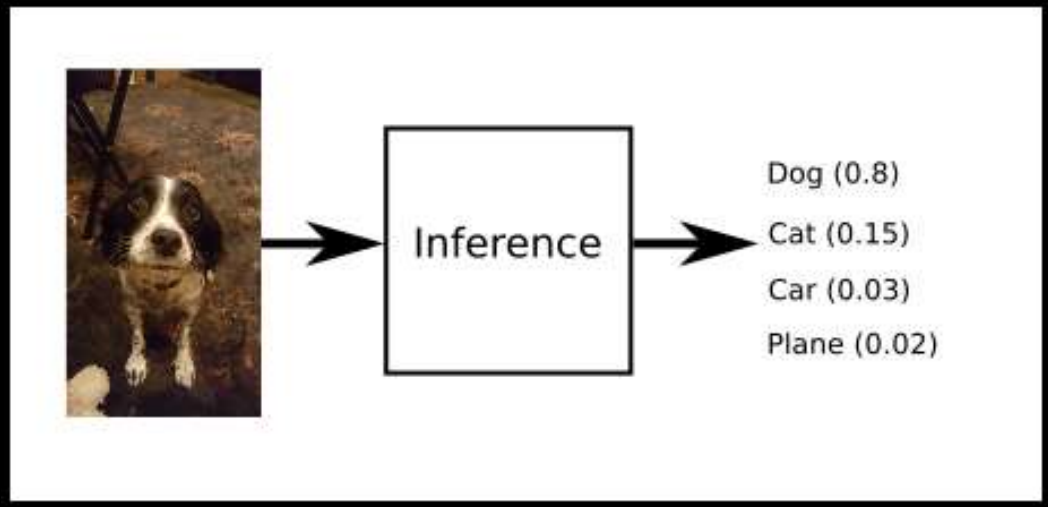
- Machine Learning is particularly interesting
  - Potential to impact nearly all software, industries
    - NLP, autonomous cars, image classification, etc.
    - “Software is eating the world, AI is eating software”
  - Compute intensive (rare)
  - Known to be extremely amenable to acceleration

WHAT IS MACHINE LEARNING?

THE ROAD TO DEEP NEURAL NETWORKS

AMD

- AI / Machine Learning:
  - Systems designed to acquire their own knowledge
  - Typically: gradient-based optimization (SGD) minimizes residual between data and model
- Rough definition of supervised learning:
  - Given a set of **data** (e.g. labelled images of dogs)
    Train a model to
    Infer properties of new data instances
    (e.g. label new images that contain dogs)
- **Training**: select parameters of model (weights)
  - Model already provided concept (feature) of dog
- **Inference**: prediction using the trained model

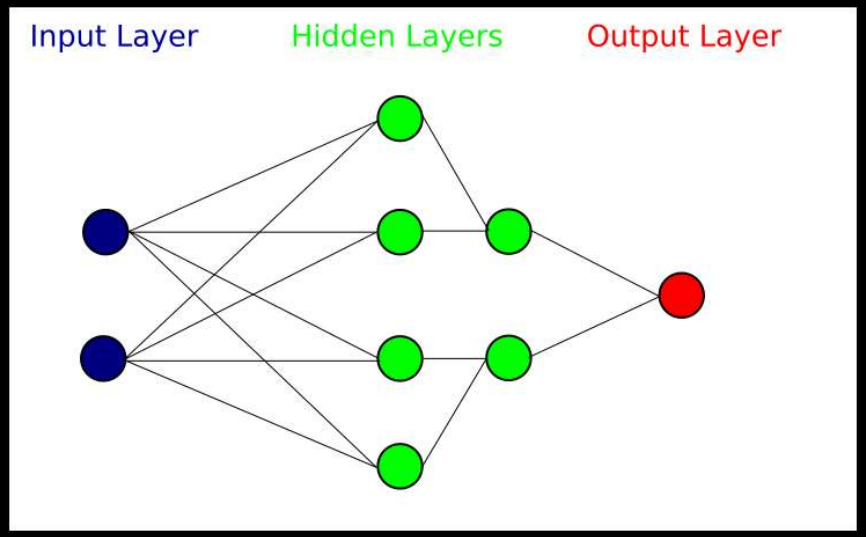


NEURAL NETWORKS

THE ROAD TO DEEP NEURAL NETWORKS

AMD

- Multilayer Perceptron / Feed-forward Neural Networks:
  - Neurons connected by synapses (weights)
  - **Input layer** (visible layer): observable variables
  - **Hidden layers**: extract increasingly abstract features from data
  - **Output**: prediction / classification of network
  - Many parameters, non-trivial to train
- Characterized by non-linear “activation function”:
  - Sigmoid, ReLU, etc.
  - Very high order basis functions
- **Hungry**
  - “Big Data”
- **Key Observation: Deeper Networks more accurate**
  - **ANN doubled in size roughly every 2.4 years**
  - **Growth driven by faster computers with larger memory**
  - GPUs achieve 95% of peak compute



WHAT ARE THE MEANS TO SUPERIOR PERFORMANCE?

REPRESENTATIVE APPROACHES

AMD

- Algorithmic:
  - FFT:
    - FFT, multiplication, iFFT
  - Strassen-Wenograd:
    - "Trades" multiplications for addition operations
- Sparse Weights:
  - Identify weights that are zero, "Prune"
  - Reduces computation, energy cost, introduces sparsity
- Reduced Precision of operations:
  - Reduced bit-width: fp16, fp8, binary, ternary
  - Reduces computation
  - Observation: inference less sensitive than training
  - Can be supported directly in hardware!

THE PROBLEM

WHAT ARE WE TRYING TO SOLVE

AMD

- No *a-priori* results capable of predicting sensitivity to precision

- Every problem must be evaluated individually
  - Requires human intervention
    - Expensive
    - Slow
    - Error-prone
    - Philosophically unappealing: autonomous algorithms should not require human intervention
  - **Mathematicians**: The solution may be unstable!
    - Could be sensitive to data, hyper-parameters, etc.
  - Objective: produce method capable of predicting numerical precision requirements for any problem
    - Heuristic?
    - Formal proof?
    - Scaling argument?

THE TEAM

IN ORDER OF IMPORTANCE

AMD

- Students (the folks actually doing the work):
  - Zhaoqi Li
  - Yu Ma
  - Catalina Vajiac
  - Yunkai Zhang
- Academic Mentor:
  - Hangjie Ji, PhD
    - Department of Mathematics, Duke University
- Industry Mentors:
  - Nicholas Malaya, PhD
  - Allen Rush, AMD Fellow
  - Alan Lee, CVP AMD, IPAM Board Member

GOALS

EASY AND MEDIUM OBJECTIVES

AMD

- **Estimate→ Easy: 100%, Medium: 50%, Hard < 32%, Stretch < 5%**
  - **Bonus:** *Where did these numbers come from?*

- **Easy:**
  - Machine Learning Frameworks (Tensorflow, Caffe, etc.) installed
  - Familiarity with core ML concepts and terminology
  - Hardware implications of mixed/reduced precision

- **Medium:**
  - Implement mixed/half/single/double precision on toy problem
  - Characterization of solution stability on toy problem

- This is research! These goals will evolve.

GOALS

HARD AND STRETCH GOALS

AMD

- **Reminder: Hard < 32%, Stretch < 5%**
- **Hard:**
  - **Stability Analysis of inference (and then, backpropagation)**
  - Can we demonstrate why inference is less sensitive to precision?
    - Likely, fewer gradients and fewer accumulated operations
    - This would be a significant result
  - **Pivot: develop statistical estimator of *a posteriori* error**
    - Likely, using Bayesian Inference
  - Scaling arguments for hyper-parameter impact on precision?
    - E.g. Neural network connectivity, depth and number of neurons
- **Stretch:**
  - Submission of results to ICML / NIPS 2018

- This is research! These goals will evolve.

FINIS

AMD

Thank you!

Questions / Comments?

nicholas.malaya@amd.com

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION