Intel Collaborative Research Institute for Computational Intelligence

Arch/ML Brainstorm Report

Shai Fine
May 24, 2016

ML/Arch Brainstorm Sessions

- **A long standing mission of ICRI-CI activity**
  - Leverage the joint ML/Arch knowledge we have in the center
- **To this end, we formed three working groups**
  - Compute model for ML
  - FPGA for ML
  - Required precision in ML
- **These were small and focused groups**
  - ML/Arch sessions revolving around a few predefined topics
    - Innovate, assess expected trends, recommend future directions, etc.
  - "Intimate" discussions
    - ~2-3 ML and 2-3 Arch lead professors are involved in each group
  - Additionally, we hope to nurture cross-domain collaborations
- **Outcomes**
  - Preliminary findings are reported in the sequel
  - Parallel breakout sessions on these topics, led by group participants

Compute Model for ML

- **Goal**
  - Discuss the need for a specific compute model for ML
  - Is there a need for new devices (such as memristor)?
    - Either as part of Von-Neumann arch. or a newly defined compute model
- **Participants**
  - HUJI: Prof. Tishby, Prof. Shalev-Shwartz
  - Technion: Prof. Ginosar, Prof. Kvatinsky
- **Status and Preliminary Observations**
  - The team decided to focused mainly on deep learning as a reference for ML compute model
    - Required compute needs, memory bandwidth, compute-memory interplay
    - The power envelop as a limiting factor (mainly for data transfer)
  - The value and properties of hierarchical framework in ML
    - A key observation – ML compute as a process of successive refinements
  - Aha Moment: Compute in memory

FPGA for ML

- **Goal**
  - Is FPGA architecture a good fit for ML?
    - What are the algorithmic/modelling implications? What are the benefits?
    - A reverse standpoint to the "Compute Model for ML" group's perspective
- **Participants**
  - Weizmann: Prof. Shamir, Prof. Nadler
  - HUJI: Prof. Fattal, Prof. Schwartz
  - Technion: Prof. Silberstein
- **Status and Preliminary Observations**
  - Reviewed the structure and characteristics of FPGAs and then project these characteristics on the ML space
  - Pipeline-like algorithms
    - Cons and pros of this notion of parallelism in the ML space
    - As a result, two members of the group started to look at concrete problems
      - Prof. Silberstein and his students - Random Forest on a HARP system
      - Prof. Fattal and Prof. Schwartz – GEMM on FPGA
  - Partial Reconfiguration
    - PR: The ability to reconfigure, off-line parts of the hardware circuitry
    - Dynamic PR: Partially re-configure while other parts are still running
    - Some parallelism was made with neuromorphic computing
  - Compute to Memory was only briefly mentioned

Required precision in ML

- **Goal**
  - Discuss the implications and the "tools" to analyze the impact of ML precision on performance (accuracy, speed, energy, memory, storage)
- **Participants**
  - Technion: Prof. Weiser, Prof. Cassuto, Prof. Manor, Prof. Crammer
  - TAU: Prof. Globerson
- **Status and Preliminary Observations**
  - Started with the definition of precision in the context of ML and then move to implications on both the HWR and on the ML outcomes
    - We've differentiated between training and scoring
    - *A key observation - Imprecise computation might be considered as a noisy computation (reminiscence of quantization), and noise occasionally helps*
    - Another interesting aspect is the tradeoff between accuracy and model complexity
    - Rigorous analysis – Reference to analysis at the "low" level of operators
  - Potential value if the hardware is allowed to make random mistakes
    - e.g. such as erroneous numeric computations, take the wrong branch
  - Breakthroughs (in ML and HWR) if precision can be altered creatively
    - e.g. self-organizing maps (Kohonen maps), Cascade of classifiers (low to high precision), hardware-based dropouts, ignoring computations that last too long

ML/Arch Breakout Sessions

- **Two parallel sessions led by group participants**

Start	End	Session	Title/Chairs		Room
13:15	14:15	Lunch + Posters			
14:15	15:15	ML/Arch Breakout Sessions	Precision in ML	Prof. Weiser, Prof. Manor	LTR
			FPGA for ML	Prof. Silberstein	PR
15:15	15:45	Break + Posters			

- **Goal**
  - Get to the next level of details
  - Discuss specific topics that the leaders chose
  - Voice your opinion

- **Hidden agenda**
  - Solicit ideas for the next Intel/Academic program