2nd AI Compute Symposium – Emerging to Pervasive (IBM, IEEE CAS, EDS)

Committee and Invited Speakers, L to R: Jin-Ping Han (IBM), Prof. Donhee Ham (Harvard/Samsung), Xin Zhang (IBM), Hsien-Hsin Sean Lee (Facebook), Matt Ziegler (IBM), Arvind Kumar (IBM), Carmen G. Almudéver (Delft University), Eduard Alarcon (UPC Barcelona Tech), Rajiv Joshi (IBM), Naresh Shanbagh (UIUC), Luis Lastras (IBM), Wen-mei Hwu (UIUC), Krishnan Kailas (IBM), Anna Topol (IBM)

http://ibm.biz/AIcomputesymposium

Dr. Rajiv Joshi's welcoming introduction

Together with the IEEE Circuits and Systems Society (CAS) and the IEEE Electron Device Society (EDS), IBM Research organized the 2nd AI Compute Symposium at the IBM T.J. Watson Research Center THINKLab in Yorktown Heights, NY, on October 17, 2019. This symposium brought together distinguished faculties, renowned thinkers, students and innovators across industry and academia together for a one-day symposium focusing on the exciting era in research addressing AI Compute from pervasive to general. The symposium consisted of three keynotes, three invited talks, a student poster session, and a panel discussion. The event was free of charge and had over 200 attendees from IBM, various companies and universities. Leadership and advancement in both pervasive and general AI domains were showcased.

Keynote talks were delivered by Dr. Luis Lastras from IBM, Prof. Wen-mei Hwu (UIUC) and Prof. Donhee Ham (Harvard/Samsung). Dr. Luis Lastras provided an exciting overview of research projects from IBM related to Natural Language Processing and its evolution. He quoted that IBM has been at the forefront of language and speech research for decades; examples are the famous research program on statistical speech processing from the 70s that led to the powerful speech recognition systems in widespread use today, the inception of BLEU - a metric widely used to measure the performance of translation systems, the well-known IBM Watson Jeopardy system that was able to defeat the world's Jeopardy champions and more recently, the demonstration of a system that can debate with a human by drawing upon new computational argumentation capabilities. In short innovations in these areas with purpose, focusing on technologies that address targeted business problems and demonstrating world class performance in strategic shared tasks are shown.

Prof. Wen-mei Hwu followed with a keynote address describing the architecture needed for AI. GPU/accelerator architectures have greatly sped up both the training and inferencing for neural-network-based machine learning models. As major industry players race to develop ambitious applications such as self-driving vehicles, unstructured data analytics, human-level interactive systems, and human intelligence augmentation, major challenges remain in computational methods as well as hardware/software infrastructures required for these applications to be effective, robust, responsive, accountable and cost-effective. These applications impose much higher levels of requirement in data storage capacity, access latency, energy efficiency, and throughput. Prof. Hwu presented a vision for building a new generation of computing components and systems for these applications.

Following the keynotes, Dr. Hsien-Hsin Sean Lee gave invited talk during the “Industry Perspectives” session about Machine Learning for Social Network Platforms. Social networks have been deeply woven into our everyday life. These Internet platforms host a plethora of real-time services to keep people connected, provide customized information to users, and preserve information transparency. Underneath their infrastructure, the adoption of Machine Learning (ML) techniques is rapidly becoming omnipresent in both datacenters and end users’ devices, steering a rich feature set to enhance the effectiveness of users’ communication and to improve the quality of online experiences. Meanwhile, to achieve these objectives, ML can consume enormous computing resources and require meticulous resource design, provisioning, and

management. Also he discussed the state-of-the-art machine learning approaches on production-scale DNN-based personalized recommendation models for content ranking and the computing challenges that lie ahead.

Next, Prof Donhee Kim (Harvard/Samsung Fellow) gave a keynote talk related to "Reconstruction of the brain". The artificial neural net made a brilliant comeback with deep learning and has since been revolutionizing a broad range of technologies in this big data age. On the other hand, deciphering the natural neuronal network of the biological brain — how it forms circuits and processes information for its higher function — is one of the most celebrated unsolved problems in all of science. He introduced on-going effort of his group to develop a semiconductor interface with biological neuronal network, which might help map and uncover its circuit and function. He further described that this study might not only contribute to the fundamental neurobiology but also help develop the next generation artificial neural net, for which several interesting possibilities were discussed.

Prof. Naresh Shanbhag from UIUC talked about "Bringing Artificial Intelligence to the Edge".
Much of AI today is deployed in the Cloud primarily due to the high complexity of machine learning algorithms. Realizing inference functionality on sensory Edge devices requires one to find ways to operate at the other edge, i.e., at the limits of energy efficiency, latency, and accuracy, in nanoscale semiconductor technologies. His talk described a Shannon-inspired model of computing (*Proceedings of the IEEE*, January 2019) to accomplish this objective. This framework comprises low signal-to-noise ratio (SNR) circuit fabrics (the channel) with engineered error statistics, coupled with efficient techniques to compensate for computational errors (encoder and decoder). A low SNR circuit fabric referred to as deep in-memory architecture (DIMA) was described. DIMA breaches the long-standing "memory wall" in von Neumann architectures by embedding analog computations in the periphery of the memory array (see https://spectrum.ieee.org/computing/hardware/to-speed-up-ai-mix-memory-and-processing) thereby achieving >100X energy-delay-product gains in laboratory prototypes over custom digital architectures implementing the same inference function. Other examples of Shannon-inspired design methods include designing deep learning systems in fixed-point, energy efficient subthreshold ECG classifier ICs, and STT-RAM based all-spin logic competitive with CMOS.

Prof. **Carmen G. Almudéver** presented current challenges in quantum computing and how the art of multi-disciplinary science is involved in it. Quantum computers promise to solve a certain set of hard problems that are intractable for even the most powerful current supercomputers. Remarkable progress has been made in recent years in quantum hardware, and quantum computation in the cloud is already a reality offering small quantum processors that are capable of handling basic quantum algorithms. Main IT companies like Google, Intel, Microsoft and IBM and numerous research groups are working on building the first universal quantum computer. Building such a quantum system requires bridging quantum algorithms and quantum processors. The talk addressed first the state of the art in quantum computing, emphasizing the main

challenges that include improvement and scalability of quantum processors, classical control electronics at (possibly) cryogenic temperatures and definition of a heterogeneous quantum computer architecture. Then, a discussion on the system architecture focusing on making quantum computing fault-tolerant and the compilation of quantum circuits was presented. In the last part of the talk, she presented her vision on how the research community could accelerate the process towards building such a scalable quantum machine, potentially through vertical cross-layer co-design structured methodologies, and possible applications, particularly quantum-enhanced Deep Learning co-processors.

The symposium also had a well-attended student poster session, where about 30 students presented compelling research spanning numerous topics in AI computing. Three best poster presentations were awarded.

Sohum Datta, Yubin Kim, and Jan Rabaey,"Statistics-inspired Architectures for the Cosine Hyper-dimensional Processor," University California, Berkeley.

Abhisek Khanna. Sourav Dutta, Jorge Gomez, Wriddhi Chakraborty, Siddharth Joshi and Suman Dutta," Spatio-Temporal Pattern Learning and classification using coupled Nano-oscillators,' University of Notre Dame.

Saruyama Pumma,Daniele Buono, Fabio Checconi, Xinnyu Que and Wu-chun Feng,"Optimizing large scale deep learning by minimizing resource contention for data processing," Virginia Tech and IBM.

The symposium closed with a panel discussion entitled "**1. What problems would we like to solve with AI that we cannot? 2. What innovations are needed to solve them?"**

Most of the invited speakers enthusiastically participated in the panel discussion, presenting their views on these two questions. One of the toughest challenges in AI is in getting enough training data to sufficiently generalize the models. A second challenge is in explain-ability of the model in inference, particularly when critical decision making is involved. Accuracy falls short of 100%, and, even more troubling, can be severely impacted by minor changes to the input. The panelists seemed to agree that solving these problems requires training concepts in addition to decisions. AI robustness, accuracy, complexity, fairness, ethics, security and privacy are all pivotal challenges. In terms of big applications, the question of whether AI can help solve big humanitarian problems such as climate change was discussed. Nearer term, the panelists considered the practicality of autonomous driving vehicles, real time wearable language translators, and understanding text and video. To solve such problems we need innovations in natural language processing, new methods in image classification and signal processing, along with high compute efficiency. With all inquisitive questions from the audience and discussions from the panel, it was apparent that AI is not just hype but offers hope.

Overall, the general consensus of attendees, speakers, and organizers was that day provided a great platform for educational forum and lively discussions related to the most current compelling topics in the computing field. Additional publications (book, journal papers, etc.) based on the symposium technical content are planned to provide educational resources for anyone interested. Although early in the stages, future events based on AI Compute are being planned by IBM and IEEE; please see the following website for updates.

Rajiv Joshi, Matt Ziegler, Arvind Kumar