Title: *Energy Efficiency and Security for Embedded AI: Challenges and Opportunities*

Abstract:

Gigantic rates of data production in the era of Big Data, Internet of Thing (IoT), and Smart Cyber Physical Systems (CPS) pose incessantly escalating demands for massive data processing, storage, and transmission while continuously interacting with the physical world under unpredictable, harsh, and energy-/power-constrained scenarios. Therefore, such systems need to support not only the high-performance capabilities under tight power/energy envelop, but also need to be intelligent/cognitive and robust. This has given rise to a new age of Machine Learning (and, in general Artificial Intelligence) at different levels of the computing stack, ranging from Edge and Fog to the Cloud. In particular, Deep Neural Networks (DNNs) have shown tremendous improvement over the past 6-8 years to achieve a significantly high accuracy for a certain set of tasks, like image classification, object detection, natural language processing, and medical data analytics. However, these DNN require highly complex computations, incurring huge processing, memory, and energy costs. To some extent, Moore's Law help by packing more transistors in the chip. However, at the same time, every new generation of device technology faces new issues and challenges in terms of energy efficiency, power density, and diverse reliability threats. These technological issues and the escalating challenges posed by the new generation of IoT and CPS systems force to rethink the computing foundations, architectures and the system software for embedded intelligence. Moreover, in the era of growing cyber-security threats, the intelligent features of a smart CPS and IoT system face new type of attacks, requiring novel design principles for enabling Robust Machine Learning.

In my research group, we have been extensively investigating the foundations for the next-generation energy-efficient and robust AI computing systems while addressing the above-mentioned challenges across the hardware and software stacks. In this talk, I will present different design challenges for building highly energy-efficient and robust machine learning systems for the Edge, covering both the efficient software and hardware designs. After presenting a quick overview of the design challenges, I will present the research roadmap and results from our Brain-Inspired Computing (BrISC) project, ranging from neural processing with specialized machine learning hardware to efficient neural architecture search algorithms, covering both fundamental and technological challenges, which enable new opportunities for improving the area, power/energy, and performance efficiency of systems by orders of magnitude. Towards the end, I will provide a quick overview of different reliability and security aspects of the machine learning systems deployed in Smart CPS and IoT, specifically at the Edge. This talk will pitch that a cross-layer design flow for machine learning/AI, that jointly leverages efficient optimizations at different software and hardware layers, is a crucial step towards enabling the wide-scale deployment of resource-constrained embedded AI systems like UAVs, autonomous vehicles, Robotics, IoT-Healthcare / Wearables, Industrial-IoT, etc.

About the Presenter:



Presenter: *Prof. Dr. Muhammad Shafique*

Affiliation:

Department of Electrical and Computer Engineering (ECE), Division of Engineering, New York University Abu Dhabi (NYUAD), UAE

ECE, Tandon School of Engineering, New York University (NYU), USA

DBLP: https://dblp.org/pers/hd/s/Shafique_0001:Muhammad

Google Scholar: https://scholar.google.com/citations?user=xSR276UAAAAJ&hl=en

Biography:

Muhammad Shafique received the Ph.D. degree in computer science from the Karlsruhe Institute of Technology (KIT), Germany, in 2011. Afterwards, he established and led a highly recognized research group at KIT for several years as well as conducted impactful R&D activities in Pakistan. Besides co-founding a technology startup in Pakistan, he was also an initiator and team lead of an ICT R&D project. He has also established strong research ties with multiple universities in Pakistan, where he is actively co-supervising various R&D activities, resulting in top-quality research outcome and scientific publications. Before, he was with Streaming Networks Pvt. Ltd. where he was involved in research and development of video coding systems several years. In Oct.2016, he joined the Institute of Computer Engineering at the Faculty of Informatics, Technische Universität Wien (TU Wien), Vienna, Austria as a Full Professor of Computer Architecture and Robust, Energy-Efficient Technologies. Since Sep.2020, he is with the Division of Engineering, New York University Abu Dhabi (NYUAD), United Arab Emirates, and is a Global Network faculty at the NYU Tandon School of Engineering, USA.

Dr. Shafique has demonstrated success in leading team-projects, meeting deadlines for demonstrations, motivating team members to peak performance levels, and completion of independent challenging tasks. His experience is corroborated by strong technical knowledge and an educational record (throughout Gold Medalist). He also possesses an in-depth understanding of various video coding standards (HEVC, H.264, MVC, MPEG-1/2/4). His research interests are in brain-inspired computing, AI & machine learning hardware and system-level design, energy-efficient systems, robust computing, hardware security, emerging technologies, FPGAs, MPSoCs, and embedded systems. His research has a special focus on cross-layer analysis, modeling, design, and optimization of computing and memory systems. The researched technologies and tools are deployed in application use cases from Internet-of-Things (IoT), smart Cyber-Physical Systems (CPS), and ICT for Development (ICT4D) domains.

Dr. Shafique has given several Keynotes, Invited Talks, and Tutorials at premier venues. He has also organized many special sessions at premier venues (like DAC, ICCAD, DATE, IOLTS, and ESWeek) and served as the Guest Editor for IEEE Design and Test Magazine (D&T) and IEEE Transactions on Sustainable Computing (T-SUSC). He has served as the TPC Chair of ISVLSI, PARMA-DITAM, RTML, ESTIMedia and LPDC, General Chair of DDECS and ESTIMedia, Track Chair at DATE, IOLTS, DSD and FDL, and PhD Forum Chair of ISVLSI. He has also served on the program committees of numerous prestigious IEEE/ACM conferences including ICCAD, ISCA, DATE, CASES, ASPDAC, and FPL. He is a senior member of the IEEE and IEEE Signal Processing Society (SPS), and a member of the ACM, SIGARCH, SIGDA, SIGBED, and HIPEAC. He holds one US patent and has (co-)authored 6 Books, 10+ Book Chapters, and over 300 papers in premier journals and conferences.

Dr. Shafique received the prestigious 2015 ACM/SIGDA Outstanding New Faculty Award (given world-wide to one person per year), the 2020 AI-2000 Chip Technology Most Influential Scholar Award, six gold medals in his educational career, and several best paper awards and nominations at prestigious conferences like CODES+ISSS, DATE, DAC and ICCAD, Best Master Thesis Award, DAC'14 Designer Track Best Poster Award, IEEE Transactions of Computer "Feature Paper of the Month" Awards, and Best Lecturer Award. His research work on aging optimization for GPUs featured as a Research Highlight in the Nature Electronics, Feb.2018 issue.