What can in-memory computing deliver, and what are the barriers?

Prof. Naveen Verma, Princeton University

Inference based on deep-learning models is being employed pervasively in applications today. In many such applications, state-of-the-art models can easily push the platforms to their limits of performance and energy efficiency. To address this, digital acceleration has been widely exploited. But, deep-learning computations exhibit critical attributes that limit the gains achievable by traditional digital acceleration. In particular, computations are dominated by high-dimensionality matrix-vector multiplications (MVMs), where the precision requirements of elements have been reducing (from FP32 a few years ago, to INT8/4/2 now and in the future). In this scenario, in-memory computing (IMC) offers distinct advantages, which have been demonstrated through recent prototypes leading to roughly 10x higher energy efficiency and area-normalized throughput, compared to optimized digital accelerators. This arises from the structural alignment of dense 2D memory arrays with the dataflow of MVMs. While digital spatial architectures (e.g., systolic arrays) also exploit this, IMC can do so more aggressively, minimizing data movement and amortizing compute into highly-efficient, highly-parallel analog operations. But, IMC also raises critical challenges, at each level (need for analog compute at circuit level, need for high bandwidth hardware infrastructure at architectural level, constrained configurability/virtualization at the software-mapping level). Recent research advances have shown remarkable promise in addressing many of these challenges, making IMC more of a reality than ever. These advances, their potential implications, and key questions remaining will be reviewed.