Accelerating, Optimizing, and Automating AI across the Stack

Dr. Kaoutar El Maghraoui, IBM, USA

Current Landscape and Future Directions Automating the decisions required for creating and setting up a deep learning (DL) pipeline has become a key direction in both research and industry to speed up the complex, slow and error-prone process of designing novel DL architectures across many modalities and tasks. Neural architecture search (NAS) has been a growing trend that aims at automating the design of neural networks that are on par or even outperform hand-designed architectures. While the holy grail for evaluating the quality of an DL model in NAS techniques has largely been focused on getting models with the highest possible accuracy, additional metrics become of paramount importance as AI models meet constrained devices (e.g., mobile and AR/VR systems) or emerging specialized hardware accelerators. Such metrics include speed or computation time, power/energy consumption, memory footprint and model size. To address these additional constraints, there is Cambrian explosion of new research directions: designing optimized deep learning hardware such as low-bit quantization and analog accelerators, efficient neural architecture search algorithms for specialized DL accelerators, and hardware (latency, energy) aware neural network architectures search targeted for constrained devices. This talk uncovers the evolving landscape of hardware-aware neural architecture search and its future directions. It also highlights IBM Research suite of techniques towards the design & build of optimized deep learning hardware as part of IBM's AI hardware Center initiative.