Precision Scaling Roadmap for AI

Abstract:

Deep learning models have been growing exponentially in parameter sizes and computational complexity, demanding significant time and energy to train and inference and generating enormous carbon emissions that make AI eco-unsustainable. As Moore's law slows down, we need a new roadmap to accelerate hardware to compute deep learning models more efficiently. By harnessing the approximate nature of deep learning models, our team has been scaling down the bit-width needed to train and inference deep learning models in the past five years, increasing the power efficiency by more than one order of magnitude. This talk will introduce several key reduced-precision techniques that enable efficient AI computations for servers and edge devices.

Bio:

Dr. Xiao Sun is a leading researcher of the Digital AI accelerator team at the IBM Thomas J. Watson Research Center. His research focuses on reduced precision training and inference of industry-scale deep learning models. Following the roadmap of bit scaling for AI in IBM, he and his team proposed the first robust 8-bit training solution, called hybrid FP8, in NeurIPS 2019 and the first-ever 4-bit training solution the following year. In ISSCC 2020, IBM integrated the Hybrid FP8 solution in a 7nm chip with 4x power efficiency than the Nvidia A100 chip. Dr. Sun earned B.S. from Peking University, M.S., and Ph.D. from Yale University, advised by Prof. T. P. Ma.