Optimizing deep learning for inference

Quantization, Pruning, and Data Reuse

In the last decade, deep learning has revolutionized various sectors, including automotive, robotics, healthcare, and security, primarily due to the increasing complexity of neural network models. However, this complexity brings significant challenges, specifically in terms of energy consumption, memory usage, and execution speed. With the shift from cloud-based to edge computing, the demand for highly efficient deep learning models has become more critical than ever. This presentation focuses on methods aimed at optimizing deep learning models for inference, with a particular focus on quantization, pruning, and data reuse.

Quantization is a technique that reduces the precision of weights and activations within a network. This reduction not only decreases the memory requirements but also speeds up computations. While quantization offers clear benefits in terms of efficiency, it is not for free: typically the accuracy of the model degrades.

Pruning removes unnecessary structures of a network, such as weights or neurons that have minimal impact on the output. This results in sparser, more efficient networks, as it reduces the memory footprint and reduces the computational load. Similar to quantization, the benefits are clear, but pruning typically also degrades the accuracy.

Data reuse addresses the significant energy costs associated with data movement in deep learning accelerators. By optimizing how data is accessed and reused across multiple levels of a memory hierarchy, this method drastically lowers energy consumption.

This presentation will explain each of these methods, discussing their rationale (why), mechanisms (how), application scenarios (where and when), and their respective advantages and constraints. Participants will leave with an understanding of the principles, benefits and limitations of each optimization technique.