LOCA SERIES: Mixed Precision Neural Networks with Second Order Taylor for the Bit Assignment

Abstract:

In this work, we present an innovative approach for neural network quantization using second-order Taylor approximations of the loss function to predict quantization error. Specifically, we calculate the Hessian of the cost function using second-order directional derivatives to model the problem as a linear programming problem. This allows us to solve it with standard solvers, finding the optimal bit assignment for each layer or group of layers. Unlike previous approaches that rely on heuristics, our method accurately computes the Hessian, considering both inter-layer and intra-layer relationships. To ensure efficiency, we compute the second-order directional derivatives, making it feasible to calculate on typical machine learning GPUs within minutes. This enables effective mixed-precision quantization of weights ranging from 2 to 8 bits. While our approach demonstrates promising capabilities, further work remains to fully explore its potential.

Speaker: Adrián Gras López

Short bio:
Adrián Gras López, is a Bachelor student of mathematics and computer science at the Polytechnic University of Catalonia (UPC). He started his studies in 2020-2021 and will graduate in the 2024-2025 academic year. Since the summer of 2023,he has been interning at the Barcelona Supercomputing Center (BSC), focusing on research in neural network quantization. During this time, he has delved into advanced techniques to improve the efficiency and performance of neural networks through quantization.

Speakers

Speaker: Adrián Gras López. Bachelor of Mathematics and Computer Science at the Polytechnic University of Catalonia (UPC).
Host: Francesc Moll. Synthesis and Physical design of ICs Group Manager, Computer Sciences, BSC. Barcelona Supercomputing Center - Centro Nacional de Supercomputación