Electronic Design Automation
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Seminar

Dyadic Neural Network Quantization

Abstract - Current low-precision quantization algorithms often have the hidden cost of conversion back and forth from floating-point to quantized integer values. This hidden cost limits the latency improvement realized by quantizing Neural Networks. To address this, we present HAWQ-V3, a novel mixed-precision integer-only quantization framework. The contributions of HAWQV3 are the following: (i) An integer-only inference where the entire computational graph is performed only with integer multiplication, addition, and bit shifting, without any floating-point operations or even integer division; (ii) A novel hardware-aware mixed-precision quantization method where the bit-precision is calculated by solving an integer linear programming problem that balances the trade-off between model perturbation and other constraints, e.g., memory footprint and latency; (iii) Direct hardware deployment and open source contribution for 4- bit uniform/mixed-precision quantization in TVM, achieving an average speedup of 1.45× for uniform 4-bit, as compared to uniform 8-bit for ResNet50 on T4 GPUs; and (iv) extensive evaluation of the proposed methods on ResNet18/50 and InceptionV3, for various model compression levels with/without mixed precision. For ResNet50, our INT8 quantization achieves an accuracy of 77.58%, which is 2.68% higher than prior integer-only work, and our mixed-precision INT4/8 quantization can reduce INT8 latency by 23% and still achieve 76.73% accuracy. Our framework and the TVM implementation have been open sourced

Contact

yu-kai.chuang@tum.de

Advisors

Yu-Kai Chuang