Dataflow-based Joint Quantization for Deep Neural Networks

Xue Geng[†⋆], Jie Fu[‡], Bin Zhao[♯], Jie Lin[†], Mohamed M. Sabry Aly[§⋆], Christopher Pal[‡], Vijay Chandrasekhar[†⋆]

[†]I²R, A⋆STAR
{geng_xue,lin-j,vijay}@i2r.a-star.edu.sg

[‡]MILA, IVADO
{jie.fu,christopher.pal}@polymtl.ca

[♯]IME, A⋆STAR
zhaobin@ime.a-star.edu.sg

[§]School of CSE, NTU
msabry@ntu.edu.sg

This paper addresses a challenging problem – how to reduce energy consumption without incurring performance drop when deploying deep neural networks (DNNs) at the inference stage[1]. In order to alleviate the computation and storage burdens, we propose a novel dataflow-based joint quantization approach with the hypothesis that a fewer number of quantization operations would incur less information loss and thus improve the final performance. It first introduces a quantization scheme with efficient bit-shifting and rounding operations to represent network parameters and activations in low precision. Given a floating-point value r, we use a quantization function, $Q(\cdot)$, to approximate it:

$$r^q = Q(r; N_r, n_{bits}) = \underbrace{min(2^{n_{bits}-1}-1, max(-2^{n_{bits}-1}, round(r \times 2^{N_r})))}_{r^I} \times 2^{-N_r} \quad (1)$$

where r^q is the quantized floating value, r^I is the integer value and N_r is the fractional bit which is the only parameter to set. Then it re-structures the network architectures to form unified modules for optimization on the quantized model. In general, four cases are considered in the same module: a) convolution layer; b) convolution layer followed by a ReLU layer; c) a residual connection with a ReLU layer; and d) a residual connection without a ReLU layer. Finally, a joint reconstruction error loss function is set up on these unified modules to do optimization on the quantized model. Extensive experiments on ImageNet and KITTI validate the effectiveness of our model. Besides, we designed and synthesized an RTL model to measure the hardware costs among various quantization methods. For each quantization operation[1], it reduces area cost by ∼15× and energy consumption by ∼9×, compared to a strong baseline.

References

[1] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, 2017.

⋆Corresponding authors.

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Project No.A1892b0026).

[1]Here, we choose to measure the energy consumption for various quantization methods on individual operations for simplicity. The overall energy consumption is in proportion to this measurement.