Ryan Lund, Anson Tsai
EE241B Project Abstract

Comparing Reduced Bitwidth Representations forMachine Learning - Quantization vs Brain Float

Abstract:

Machine learning (ML) has rapidly become an essential component of modern commercial products. As such, researchers have sought to optimize both software and hardware in order to support faster and more energy efficient training and inference, both on the warehouse scale and for mobile devices. A recent area of focus is to use reduced bit-width numerical representation for weights and activations. One technique to reduce bitwidth is quantization, or using narrow integers along with a scaling factor in the place of a wider float value. An alternative technique is to use a novel numerical representation, such as the 16 bit bFloat format (bf16). With 1 sign, 8 exponent, and 7 explicit mantissa bits, bf16 is able to preserve the approximate dynamic range and model accuracy of a 32-bit floating point (fp32) network while consuming half the space. In this project, we will compare the impacts on accuracy, performance, and resource utilization of these reduced bitwidth methods by running a testbench of ML workloads on a Gemmini generated systolic array coupled with Rocket Core in-order and BOOM out-of-order processors.

References:

G. Tagliavini, S. Mach, D. Rossi, A. Marongiu, and L. Benini. A transprecision floating-point platform for ultra-low power computing. arXiv:1711.10374 [cs.DC], 2017.

H. Genc, A. Haj-Ali, V. Iyer, A. Amid, H. Mao, J. Wright, C. Schmidt, J. Zhao, A. Ou, M. Banister, Y. S. Shao, B. Nikoli´c, I. Stoica, and K. Asanovi´c. Gemmini: An agile systolic array generator enabling systematic evaluations of deep-learning architectures. arXiv:1911.09925 [cs.DC], 2019.

J. Y. F. Tong, D. Nagle and R. A. Rutenbar, "Reducing power by optimizing the necessary precision/range of floating-point arithmetic," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 8, no. 3, pp. 273-286, June 2000.

N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan. Training deep neural networks with 8-bit floating point numbers. arXiv:1812.08011 [cs.DC], 2018.

S. Wang, and P. Kanwar. “BFloat16: The Secret to High Performance on Cloud TPUs” Google, Google, 23 Aug. 2019, cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus.