Ryan Lund, Anson Tsai
EE241B Project Abstract

Benefits and Tradeoffs of Utilizing bfloat16 in Systolic Array ML Training/Inference

Abstract:
With machine learning (ML) becoming an essential component of modern commercial products, researchers have sought to optimize both software and hardware in order to support faster and more energy efficient inference. Recently, a new float format named bfloat16 was adopted for training/inference optimization. By having 1 sign, 8 exponent, and 7 explicit mantissa bits, bfloat16 is able to preserve the approximate dynamic range of float32 while having the area footprint of a float16. Moreover, research has shown that the truncation of mantissa bits have minimal effect on the accuracy of models. This has many cost saving implications, as models now use less space and more data can be transferred/computed per cycle. In our research project, we will evaluate the benefits and tradeoffs of utilizing bfloat16 in ML training/inference workloads using systolic arrays. To gather data, we plan on using Gemmini, a systolic array generator, coupled with Rocket Chip and the BOOM out-of-order processor to run various workloads that compare the accuracy, performance, and resource utilization of using bfloat16 versus other number formats such as float32 and integer8.

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.