Abstract

The IEEE 754 floating point standard has long been the go-to standard for high-precision mathematical computations. Recent developments in deep learning have, however, lead researchers to search for alternatives to the standard that would best complement the computations required in deep learning. One current popular alternative is Google's Bfloat16 used in their Tensor Processing Unit (TPU), which increases the dynamic range of IEEE 754 FP16 by using an 8-bit wide exponent, similar to the IEEE FP32 standard, while compromising on precision with the reduced mantissa.

In this project, we would like to examine an alternative number system to floating point, introduced in 2017, known as Posits. Posits have a wider dynamic range compared to floating point, while maintaining similar precision at values close to zero. These characteristics have prompted researchers to examine Posits as a number representation for deep learning applications. We would like to investigate the hardware implications of the Posit representation in fused multiply-add operations used frequently in deep learning matrix multipliers and compare them to those involved in FP16 and Bfloat16 representations. If time permits, we would also like to look into the actual implementation of the FMA unit in a systolic array and compare the situations under which each number representation excels.

References

[1] P. Lindstrom, S. Lloyd, and J. Hittinger, "Universal coding of the reals: Alternatives to ieee floating point", in *Proceedings of the Conference for Next Generation Arithmetic*, ser. CoNGA '18, Singapore, Singapore: Association for Computing Machinery, 2018, ISBN: 9781450364140. DOI: 10.1145/3190339.3190344. [Online]. Available: https://doi.org/10.1145/3190339.3190344.

[2] R. Chaurasiya, J. Gustafson, R. Shrestha, J. Neudorfer, S. Nambiar, K. Niyogi, F. Merchant, and R. Leupers, "Parameterized posit arithmetic hardware generator", in *2018 IEEE 36th International Conference on Computer Design (ICCD)*, Oct. 2018, pp. 334–341. DOI: 10.1109/ICCD.2018.00057.

[3] F. de Dinechin, L. Forget, J.-M. Muller, and Y. Uguen, "Posits: The good, the bad and the ugly", in *Proceedings of the Conference for Next Generation Arithmetic 2019*, ser. CoNGA'19, Singapore, Singapore: Association for Computing Machinery, 2019, ISBN: 9781450371391. DOI: 10.1145/3316279.3316285. [Online]. Available: https://doi.org/10.1145/3316279.3316285.

[4] J. Johnson, *Rethinking floating point for deep learning*, 2018. arXiv: 1811.01721 [cs.NA].

[5] Gustafson and Yonemoto, "Beating floating point at its own game: Posit arithmetic", *Supercomput. Front. Innov.: Int. J.*, vol. 4, no. 2, pp. 71–86, Jun. 2017, ISSN: 2409-6008. DOI: 10.14529/jsfi170206. [Online]. Available: https://doi.org/10.14529/jsfi170206.