Selected References on Binary8 Floating-Point Formats

FP8 versus INT8 for efficient deep learning inference

> Van Ballen, et al,. 2023, "FP8 versus INT8 for efficient deep learning inference", https://arxiv.org/pdf/2303.17951.pdf

Recently, the idea of using FP8 as a number format for neural network training has been floating around the deep learning world. Nvidia has announced the FP8 format for their transformer engine software for the new Hopper architecture GPUs, and an IEEE consortium is investigating the standardization of the FP8 format for training deep learning networks in the cloud. Given that most training is currently conducted with entire networks in FP32, or sometimes FP16 with mixed precision, the step to having some parts of a network run in FP8 with 8-bit weights is an appealing potential speed-up for the generally costly and time-intensive training procedures in deep learning. (from the abstract)

> Section 4.1, Figure 5
>
> … We can look at a few simple distributions to see what this difference in the distribution of points means. We consider the mean-squared error of these distributions, as this has been shown to correlate strongly, both mathematically and practically, with the effect of noise on neural networks (Nagel et al. (2020a)). In Figure 5, we can see that the INT format is the best for a uniform distribution. This is natural, as the uniformly distributed grid points match the real-valued uniform distribution. For the Gaussian distribution in the middle column, [binary8p6] is best [of binary8p4, binary8p5, binary8p6] … [binary8p6 is best for the Uniform distribution, binary8p4 is best for the Student-T distribution (v=3)]

Efficient Post-training Quantization with FP8 Formats

> Shen, H., et. al., 2023, "Efficient Post-training Quantization with FP8 Formats.", https://arxiv.org/2309.14592

Recent advances in deep learning methods such as LLMs and Diffusion models have created a need for improved quantization methods that can meet the computational demands of these modern architectures while maintaining accuracy. Towards this goal, we study the advantages of FP8 data formats for post-training quantization across 75 unique network architectures covering a wide range of tasks, including machine translation, language modeling, text generation, image classification, generation, and segmentation. We examine three different FP8 representations (E5M2, E4M3, and E3M4) to study the effects of varying degrees of trade-off between dynamic range and precision on model accuracy. Based on our extensive study, we developed a quantization workflow that generalizes across different network architectures. Our empirical results show that FP8 formats outperform INT8 in multiple aspects, including workload coverage (92.64% vs. 65.87%), model accuracy and suitability for a broader range of operations. Furthermore, our findings suggest that E4M3 is better suited for NLP models, whereas E3M4 performs marginally better than E4M3 on computer vision tasks. (from the abstract)

With Shared Microexponents, A Little Shifting Goes a Long Way

Rouhani, B., et al,. 2023, "With Shared Microexponents, A Little Shifting Goes a Long Way.", https://arxiv.org/pdf/2302.08007.pdf

This paper introduces Block Data Representations (BDR), a framework for exploring and evaluating a wide spectrum of narrow-precision formats for deep learning. It enables comparison of popular quantization standards, and through BDR, new formats based on shared microexponents (MX) are identified, which outperform other state-of-the-art quantization approaches, including narrow-precision floating-point and block floating-point. … (from the abstract)

Figure 7: QSNR[1] vs. the area-memory efficiency product for different .. configurations. We focus our search space on symmetric dot product units. The area is normalized to a 64-element FP8 dot product and the memory efficiency is the inverse of the packing efficiency of a 256-element tile into [64B].

binary8p5 is 1st in both QSNR and memory efficiency	QSNR = 35, eff = 1.00 [2]
binary8p4 is 2nd in both QSNR and memory efficiency	QSNR = 30, eff = 0.85
binary8p3 is 3rd in both QSNR and memory efficiency	QSNR = 25, eff = 0.65

[1] .. we use Quantization Signal to Noise Ratio (QSNR) as a measure of numerical fidelity .. (pg 4)

[2] The values given for QSNR and memory efficiency (eff) are estimated from Figure 7.

8-bit Numerical Formats for Deep Neural Networks

Noune, B., et al,. 2022, "8-bit Numerical Formats for Deep Neural Networks.", http://arxiv.org/abs/2206.02915

Given the current trend of increasing size and complexity of machine learning architectures, it has become of critical importance to identify new approaches to improve the computational efficiency of model training. In this context, we address the advantages of floating-point .., and present .. the use of 8-bit floating-point number formats for activations, weights, and gradients for both training and inference. We explore the effect of different bit-widths for exponents and significands and different exponent biases. … (from the abstract)

Table 1 (pg 3):

format	precision	significand trailing bits	exponent bits	Dynamic Range [DR] (dB)	Signal to Noise Ratio [SNR] (dB)
binary8p3	3	2	5	197.5	25.5
binary8p4	4	3	4	107.8	31.5
binary8p5	5	4	3	66.0	37.5

ZeroQuant-FP: A Leap Forward in LLMs Post-Training .. Using Floating-Point Formats

> Wu, X., et. al., 2023, "ZeroQuant-FP: A Leap Forward in LLMs Post-Training W4A8 Quantization Using Floating-Point Formats.", https://arxiv.org/abs/2307.09782

In the complex domain of large language models (LLMs), striking a balance between computational efficiency and maintaining model quality is a formidable challenge. Navigating the inherent limitations of uniform quantization, particularly when dealing with outliers, and motivated by the launch of NVIDIA's H100 hardware, this study delves into the viability of floating-point (FP) quantization, particularly focusing on FP8 and FP4, as a potential solution. (from the abstract)

We find that .. the configuration E4M3 [outperforms] E5M2 … (pg 6)

Exploring the Potential of Flexible 8-bit Format: Design and Algorithm

> Zhang, Z., et al,. 2023, "Exploring the Potential of Flexible 8-bit Format: Design and Algorithm.", https://arxiv.org/pdf/2310.13513.pdf

Neural network quantization is widely used to reduce model inference complexity in real-world deployments. However, traditional integer quantization suffers from accuracy degradation when adapting to various dynamic ranges. Recent research has focused on a new 8-bit format, FP8, with hardware support for both training and inference of neural networks but lacks guidance for hardware design. (from the abstract)

> Most of the FP8 formats are designed according to the IEEE-754 standard. [.. there are also formats, such as the NIA(Nvidia-Intel-Arm) designed format (Micikevicius et al. 2022), which .. expand the range..] . Our research indicates that the primary challenge with FP8 quantization is often insufficient precision in characterizing values close to 0, rather than a limitation of the dynamic range .. .
>
> For instance, the E2M5 [binary8p6] format has a maximum absolute normal value in IEEE-754 format of S.10.111112 =1.96875 ∗ 21 = 3.9375 and a minimum absolute normal value of S.01.00000 = 20 = 1. Consequently, any value from -1 to 1 will be truncated to 0. Moreover, it is arduous to scale the magnitude of original data to the range of [1, 3.9375], leading to inadequate quantization accuracy, particularly when using the FP8 format with lower exponents. Therefore, we introduced subnormal values into our number system to increase the representation capability of values in the range of (-1, 1). This allowed E2M5 to represent the minimum absolute value of S.00.000012 = 2−5 = 0.03125 … (section 4.1)