1-bit LLMs

23rd March 2024

LLMs

1. Expensive (high inference costs and energy consumption)
2. As size grows → More hardware required for inference (bottleneck)
3. Environmental and economic impact due to high energy consumption
4. Transferring model parameters from DRAM to the memory of an on-chip accelerator (e.g., SRAM) can be expensive during inference.
5. Is it possible? mathematically..

Quantization

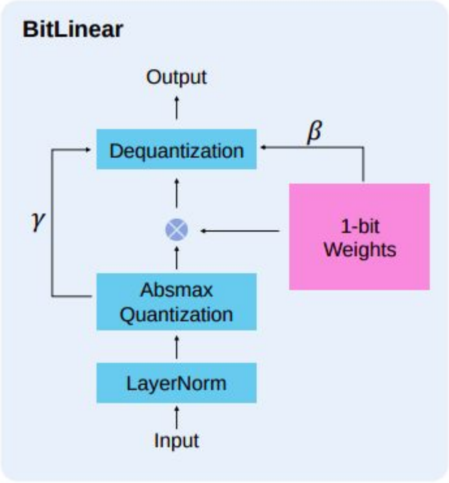
1. Post training quantization ~ leads to decrease in accuracy.
6. Quantization aware training ~ model difficult to converge with lower precision.
   a. (kqv, feed forward in transformers)

Binarization

1. Extreme case of Quantization
2. Energy and compute efficient
3. Needs to scale with size of the model (in same manner as full-precision Transformers).

BitNet: Scaling 1-bit Transformers for Large Language Models (2023)

1. Matrix multiplication replaced by BitLinear
   a. model weights 1-bit are -1/1 (signum fn) Wf = Sign(W − α)
   b. scaling factor β is used after binarization to reduce the l2 error between the real-valued and the binarized weights.
   c. other components are 8-bit
      not that compute intensive
      high precision is needed for LLMs to sample
   d. absmax quantization of the input→ [-1, 1]
   e. LayerNorm ensures the mag of var(y) in quantized version is of the same mag as var(y) with full precision.

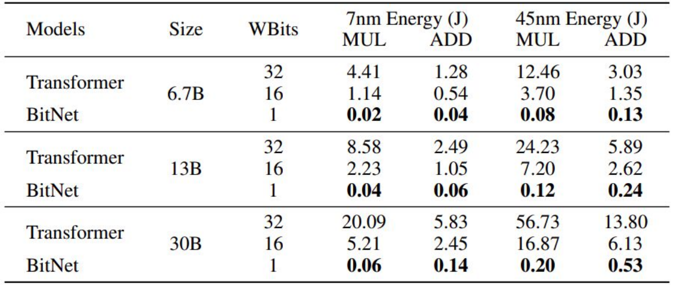


2. Operations
   a. The matrix multiplication is performed between the 1-bit weights and the quantized activations.
   b. The output activations are rescaled with {β, γ} to dequantize them to the original precision.

BitNet: Scaling 1-bit Transformers for Large Language Models (2023)

1. Model parallelism trick
    1. Pre-requisite is tensors are independent along the partition dimension
    2. Quantization params {α, β, γ} need to be estimated for from the whole tensors.
    3. Weights and activations into groups and then independently estimate each group’s parameters.

2. Training tricks
    1. Bypasses the nondifferentiable functions
    2. Mixed precision training for gradients and weights
    3. Large learning rate needed since small gradients dont change weights from -1/1.

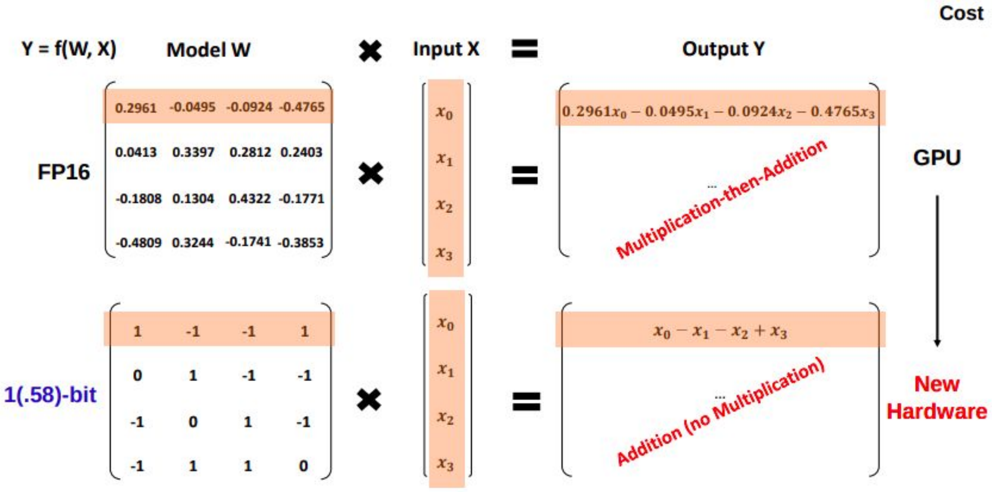


Models	Size	WBits	7nm Energy (J)		45nm Energy (J)	
			MUL	ADD	MUL	ADD
Transformer	6.7B	32	4.41	1.28	12.46	3.03
		16	1.14	0.54	3.70	1.35
BitNet		1	**0.02**	**0.04**	**0.08**	**0.13**
Transformer	13B	32	8.58	2.49	24.23	5.89
		16	2.23	1.05	7.20	2.62
BitNet		1	**0.04**	**0.06**	**0.12**	**0.24**
Transformer	30B	32	20.09	5.83	56.73	13.80
		16	5.21	2.45	16.87	6.13
BitNet		1	**0.06**	**0.14**	**0.20**	**0.53**

Table 1: Energy consumption of BitNet and Transformer varying different model size. Results are reported with 512 as input length.

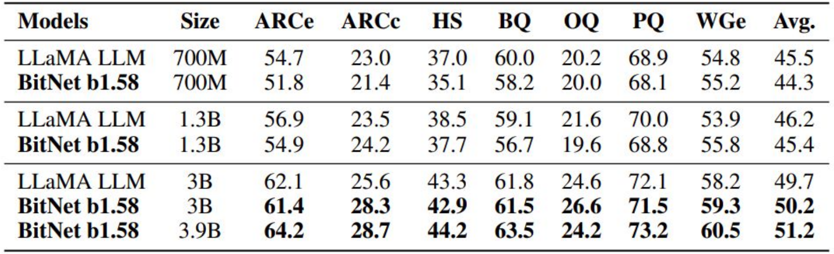
The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits (2024)

1. Model: BitNet b1.58
   a. model weights {-1, 0, 1}
   b. same energy consumption as 1-bit BitNet
2. Advantages
   a. stronger due to its explicit support for feature filtering (0)
   b. BitNet b1.58 can match FP16.
3. Results
   a. matches FP16 > 3B model
   b. Requires 3-8 times < memory.
4. Outlook
   a. Long Sequence in LLMs
   b. LLMs on Edge and Mobile
   c. New Hardware for 1-bit LLMs



$$\begin{pmatrix} 0.2961 & -0.0495 & -0.0924 & -0.4765 \\ 0.0413 & 0.3397 & 0.2812 & 0.2403 \\ -0.1808 & 0.1304 & 0.4322 & -0.1771 \\ -0.4809 & 0.3244 & -0.1741 & -0.3853 \end{pmatrix} \times \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.2961x_0 - 0.0495x_1 - 0.0924x_2 - 0.4765x_3 \\ \dots \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & 0 & 1 & -1 \\ -1 & 1 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_0 - x_1 - x_2 + x_3 \\ \dots \end{pmatrix}$$



Models	Size	ARCe	ARCc	HS	BQ	OQ	PQ	WGe	Avg.
LLaMA LLM	700M	54.7	23.0	37.0	60.0	20.2	68.9	54.8	45.5
BitNet b1.58	700M	51.8	21.4	35.1	58.2	20.0	68.1	55.2	44.3
LLaMA LLM	1.3B	56.9	23.5	38.5	59.1	21.6	70.0	53.9	46.2
BitNet b1.58	1.3B	54.9	24.2	37.7	56.7	19.6	68.8	55.8	45.4
LLaMA LLM	3B	62.1	25.6	43.3	61.8	24.6	72.1	58.2	49.7
BitNet b1.58	3B	**61.4**	**28.3**	**42.9**	**61.5**	**26.6**	**71.5**	**59.3**	**50.2**
BitNet b1.58	3.9B	**64.2**	**28.7**	**44.2**	**63.5**	**24.2**	**73.2**	**60.5**	**51.2**

Table 2: Zero-shot accuracy of BitNet b1.58 and LLaMA LLM on the end tasks.