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

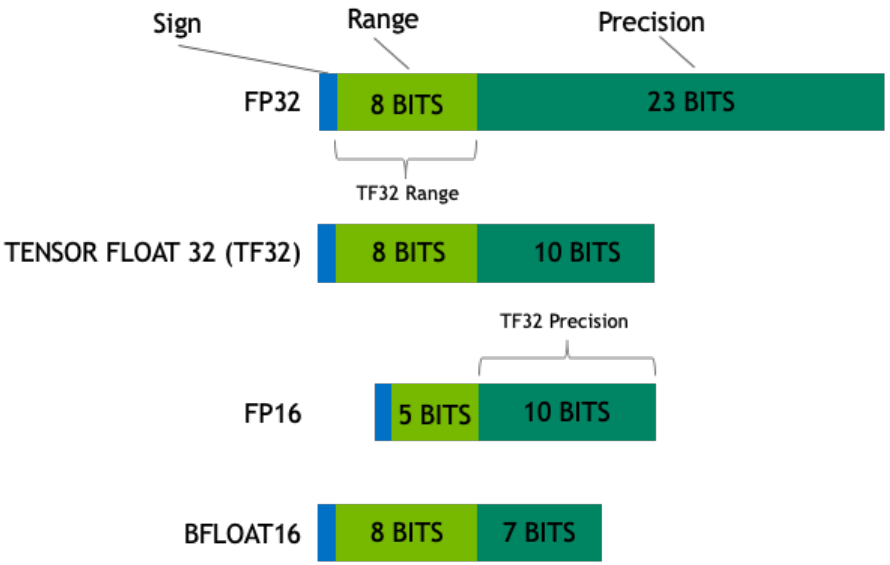
Shuming Ma* Hongyu Wang* Lingxiao Ma Lei Wang Wenhui Wang
Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei◇
https://aka.ms/GeneralAI

HCCLab


Junghwan Kim

Background

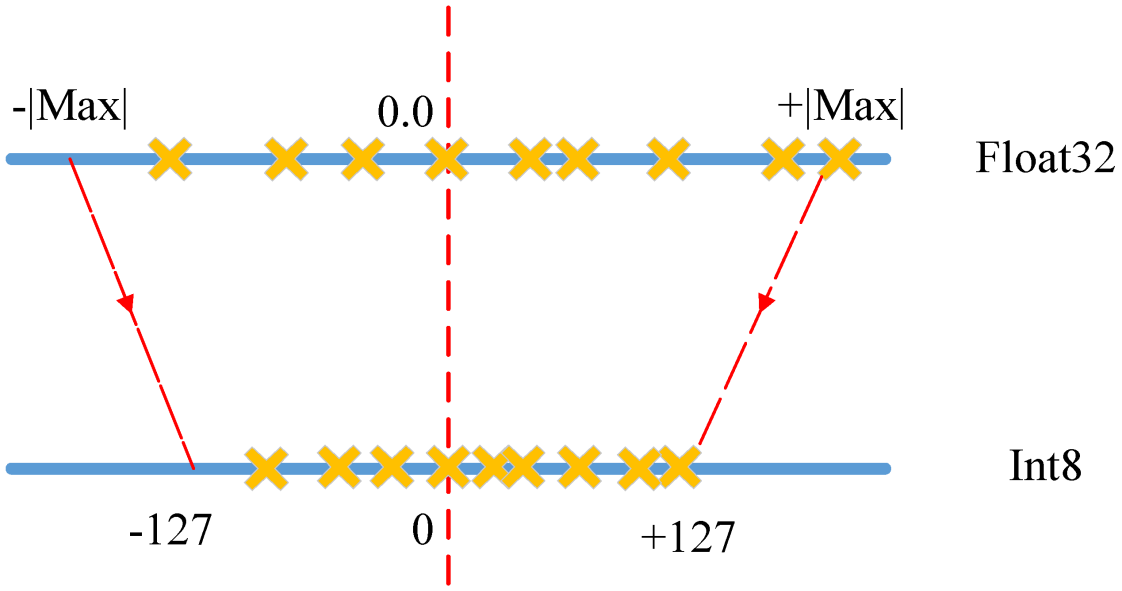
- Increasing Size Posed Environmental and Economic Impact due to High Enery Consumptions
  - Vanilla LLMs are in 16-bit floating values(FP16 or BF16)
  - Bulk of Any LLM is matrix multiplication
  - Rise of Quantization



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

Background

- Increasing Size Posed Environmental and Economic Impact due to High Enery Consumptions
  - Vanilla LLMs are in 16-bit floating values(FP16 or BF16)
  - Bulk of Any LLM is matrix multiplication
  - Rise of Quantization



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

Background - BitNet

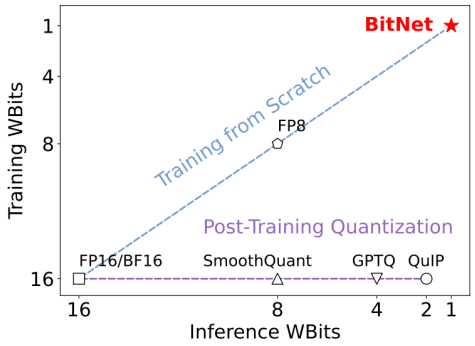
BitNet: Scaling 1-bit Transformers for Large Language Models

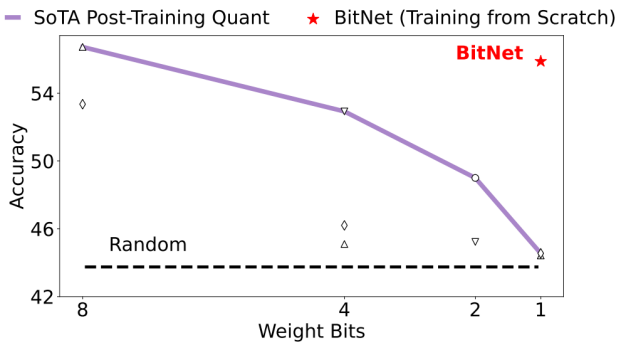
Hongyu Wang[*†‡] **Shuming Ma**[*†] **Li Dong**[†] **Shaohan Huang**[†]
Huaijie Wang[§] **Lingxiao Ma**[†] **Fan Yang**[†] **Ruiping Wang**[‡] **Yi Wu**[§] **Furu Wei**[†◇]
[†] Microsoft Research [‡] University of Chinese Academy of Sciences [§] Tsinghua University
https://aka.ms/GeneralAI

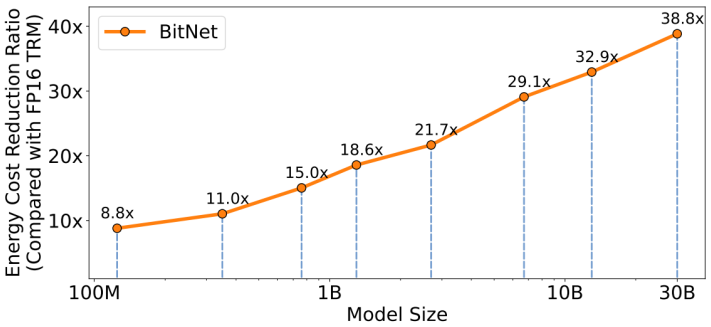
> I don't think there's anything unique about human intelligence. All the neurons in the brain that make up perceptions and emotions operate in a binary fashion.
>
> *William Henry Gates III*

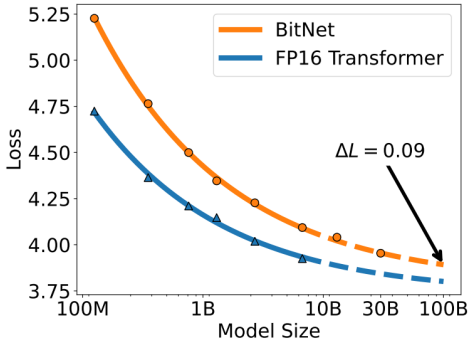
The Era of 1-bit LLMs : All Large Language Models are in 1.58 Bits(2024)

Background - BitNet





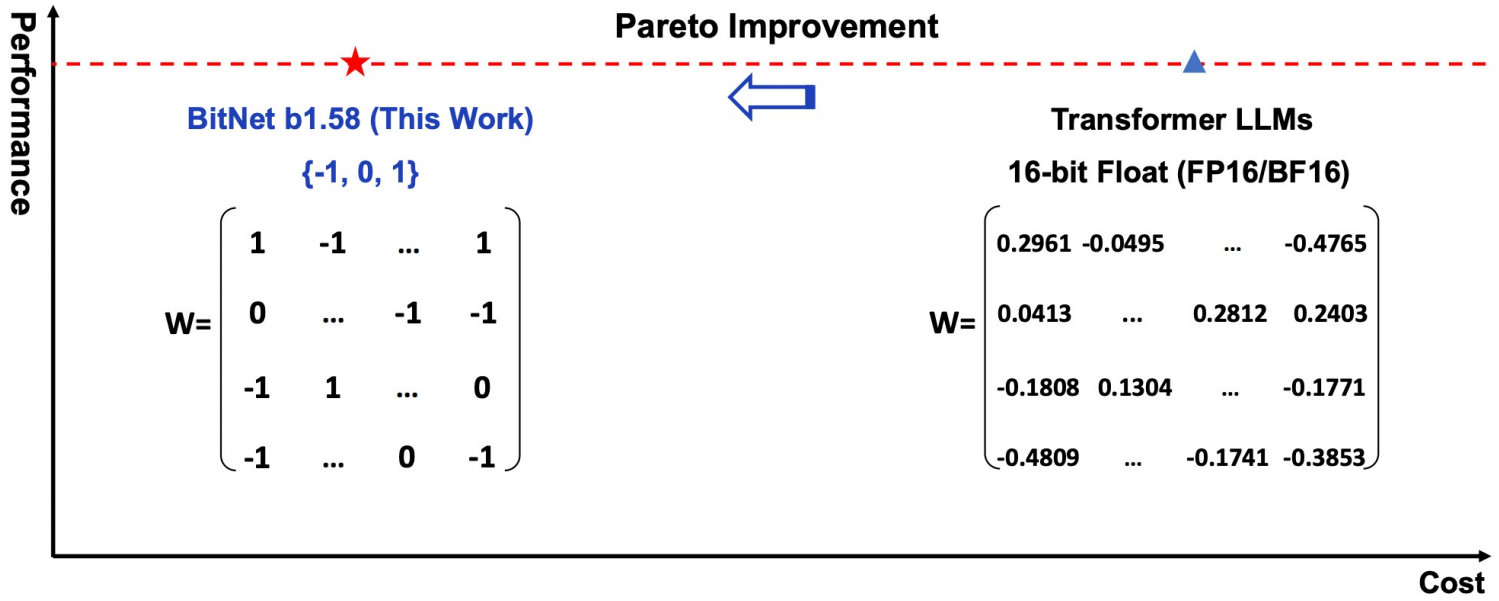




- Model size scales **UP**, the cost saving scales **UP**
- **BUT**, Loss convergence problem..

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

Method



- BitNet : {-1, 1}
- BitNet b1.58 : {-1, 0, 1}

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

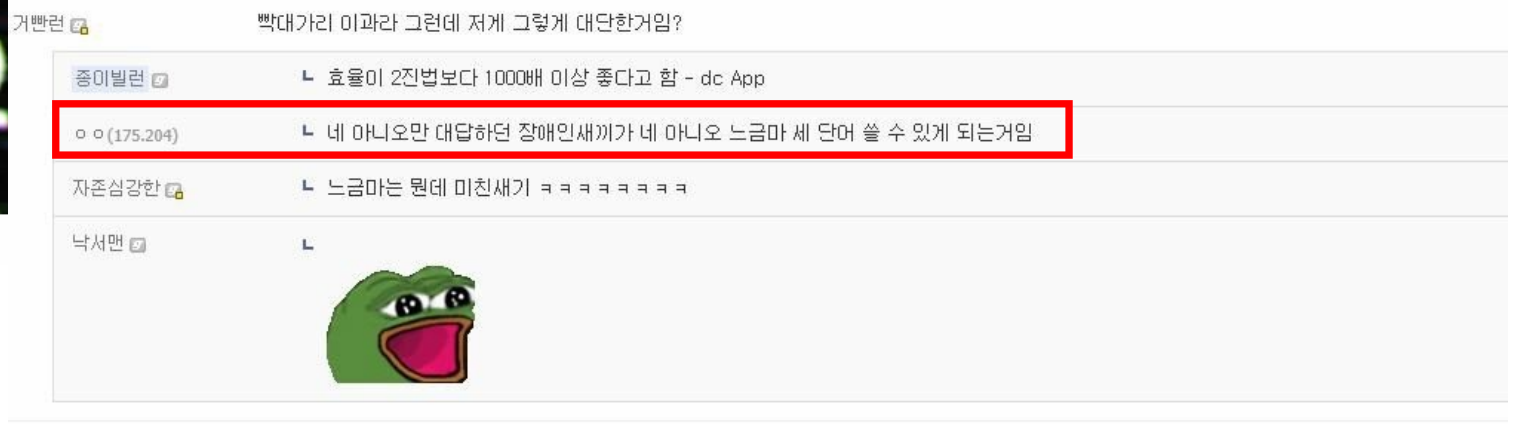
Method

미국과 중국, 한국과 일본 등 반도체를 둘러싼 주요국들의 경쟁이 치열한 가운데 국내 연구진의 '기술 낭보'가 전해졌다.

17일 삼성전자에 따르면 울산과학기술원(UNIST) 전기·전자·컴퓨터공학부 김경록 교수 연구팀은 초절전 '3진법 금속-산화막-반도체(Ternary Metal-Oxide-Semiconductor)'를 세계 최초로 큰 면적의 실리콘 웨이퍼에서 구현하는 데 성공했다고 밝혔다. 연구 결과는 지난 15일(영국 현지시각) 세계적인 학술지 『네이처 일렉트로닉스(Nature Electronics)』에 발표됐다. 삼성전자는 이번 연구를 2017년 9월 삼성 미래기술육성사업 지정 테마로 선정해 지원해 왔다.

3진법 탄생함

디지몬좌 선견지명
- dc official App



- BitNet : {-1, 1}
- BitNet b1.58 : {-1, 0, 1}
  - Convergence - Faster
  - Quantization - Memory
  - Inference Speed - Faster

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

Method

Quantization Function. To constrain the weights to -1, 0, or +1, we adopt an *absmean* quantization function. It first scales the weight matrix by its average absolute value, and then round each value to the nearest integer among {-1, 0, +1}:

$$\widetilde{W} = \text{RoundClip}(\frac{W}{\gamma + \epsilon}, -1, 1), \tag{1}$$

$$\text{RoundClip}(x, a, b) = \max(a, \min(b, \text{round}(x))), \tag{2}$$

$$\gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|. \tag{3}$$

- Code is Not opened yet..
  - RoundClip : X값을 -1와 1 사이에서 제일 근접한 정수로 반올림
  - -1보다 작거나 1보다 크면 -1 또는 1로 Clipping

Result

Models	Size	Memory (GB)↓	Latency (ms)↓	PPL↓
LLaMA LLM	700M	2.08 (1.00x)	1.18 (1.00x)	12.33
BitNet b1.58	700M	0.80 (2.60x)	0.96 (1.23x)	12.87
LLaMA LLM	1.3B	3.34 (1.00x)	1.62 (1.00x)	11.25
BitNet b1.58	1.3B	1.14 (2.93x)	0.97 (1.67x)	11.29
LLaMA LLM	3B	7.89 (1.00x)	5.07 (1.00x)	10.04
BitNet b1.58	3B	**2.22 (3.55x)**	**1.87 (2.71x)**	**9.91**
BitNet b1.58	3.9B	**2.38 (3.32x)**	**2.11 (2.40x)**	**9.62**

Table 1: Perplexity as well as the cost of BitNet b1.58 and LLaMA LLM.

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.

Result

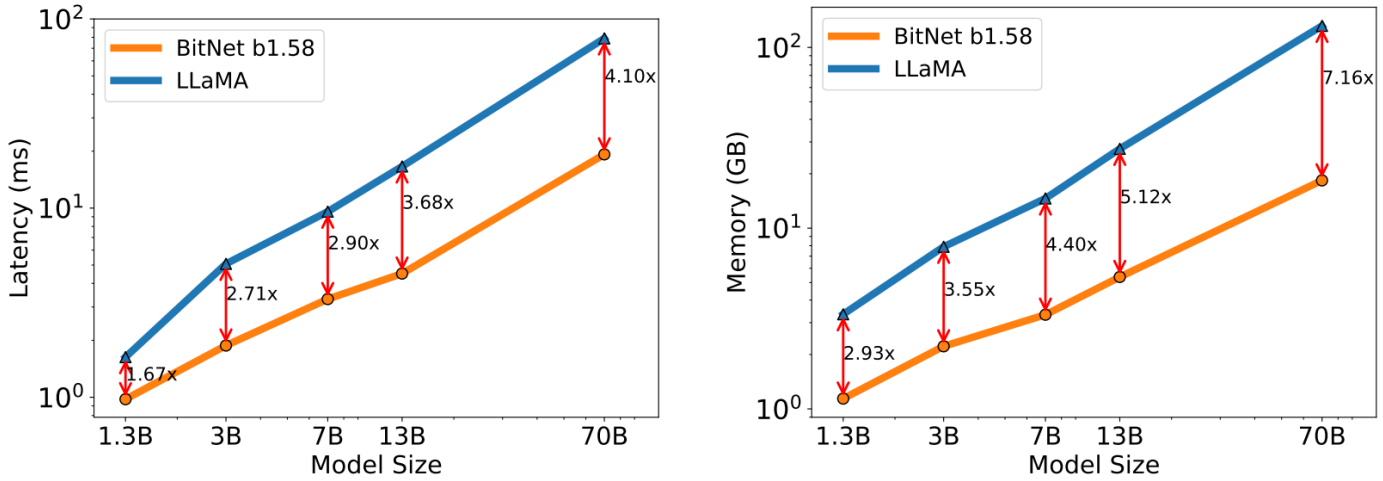


Figure 2: Decoding latency (Left) and memory consumption (Right) of BitNet b1.58 varying the model size.

Result

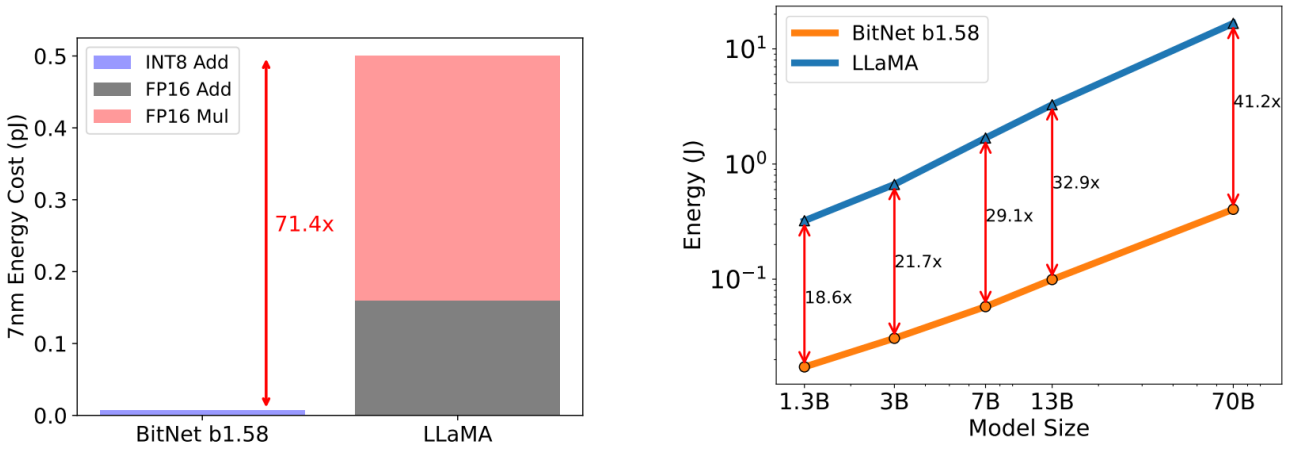


Figure 3: Energy consumption of BitNet b1.58 compared to LLaMA LLM at 7nm process nodes. On the left is the components of arithmetic operations energy. On the right is the end-to-end energy cost across different model sizes.

- BitNet b1.58 is enabling new scaling law

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

Limitation & Future Works

- 1-bit Mixture of Experts(MoE) LLMs
  - MoE는 LLM에 대한 비용 효율적인 칩 간 통신 시스템
- Native Support of Long Sequence in LLMs
  - 같은 자원을 가진 상황에서 Context 길이 두배 이상 늘릴 수 있음
- LLMs on Edge and Mobile
  - 오히려 CPU 친화적, 감소된 메모리 및 에너지 소비
- New Hardware for 1-bit LLMs
  - LLM을 위한 하드웨어(LPU) 구축 과정에서 덧셈 기반 새로운 구조

Conclusion

ybelkada 8 days ago

Very nice paper that introduces a new paradigm for LLM quantization (ternary weights for linear layers {-1, 0, 1} resulting in removing the need of having multiplications in matmul + int8 activations)
It seems that method cannot be used as a post-training quantization method, but rather train a 1.5-bit model from scratch. I believe the code will be shared here: https://github.com/microsoft/unilm/tree/master/bitnet - would be curious to see if the authors will share the quantized models on the Hub!
I also wonder if the lm_head is also quantized, as not quantizing the lm head helps for preserving good generation quality for quantized language models

6 replies · ❤️ 45 🤗 2 +

shumingma Paper author 8 days ago

We would definitely be happy to open-source the models for future research. Please stay tuned!

The lm_head is not quantized because the language models have to use high-precision
probabilities to perform sampling, and it only takes a very small proportion of the cost especially when the model is large.

👍 97 ❤️ 41 🤝 12 +

Expand 5 replies

Reply in thread

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

Thanks😍

HCC LAB SNU