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

2024.02.27

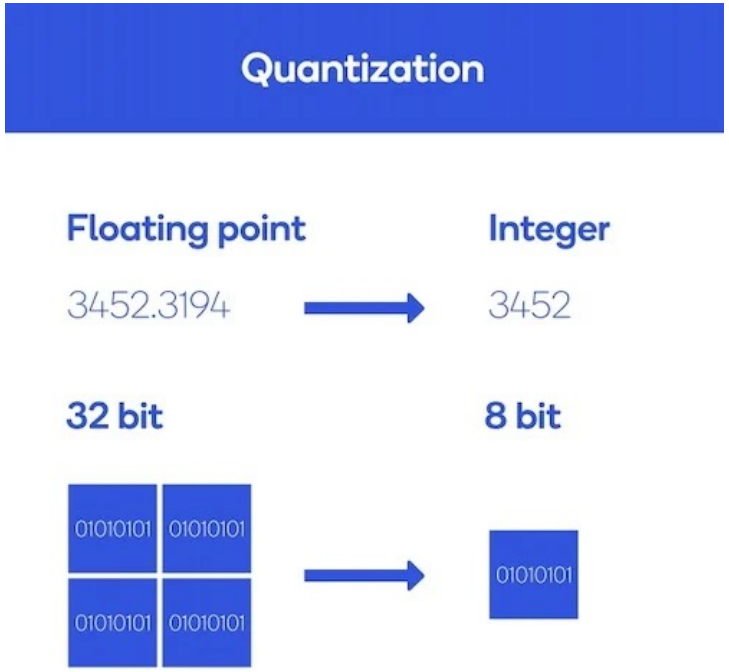
발표자 : 김산

Introduction

양자화?

신경망의 가중치와 활성 함수의 출력값을 더 작은 bit로 변환하는 것.

⟶ 모델의 성능을 유지하면서 크기는 줄이고 계산 비용을 낮추는 것

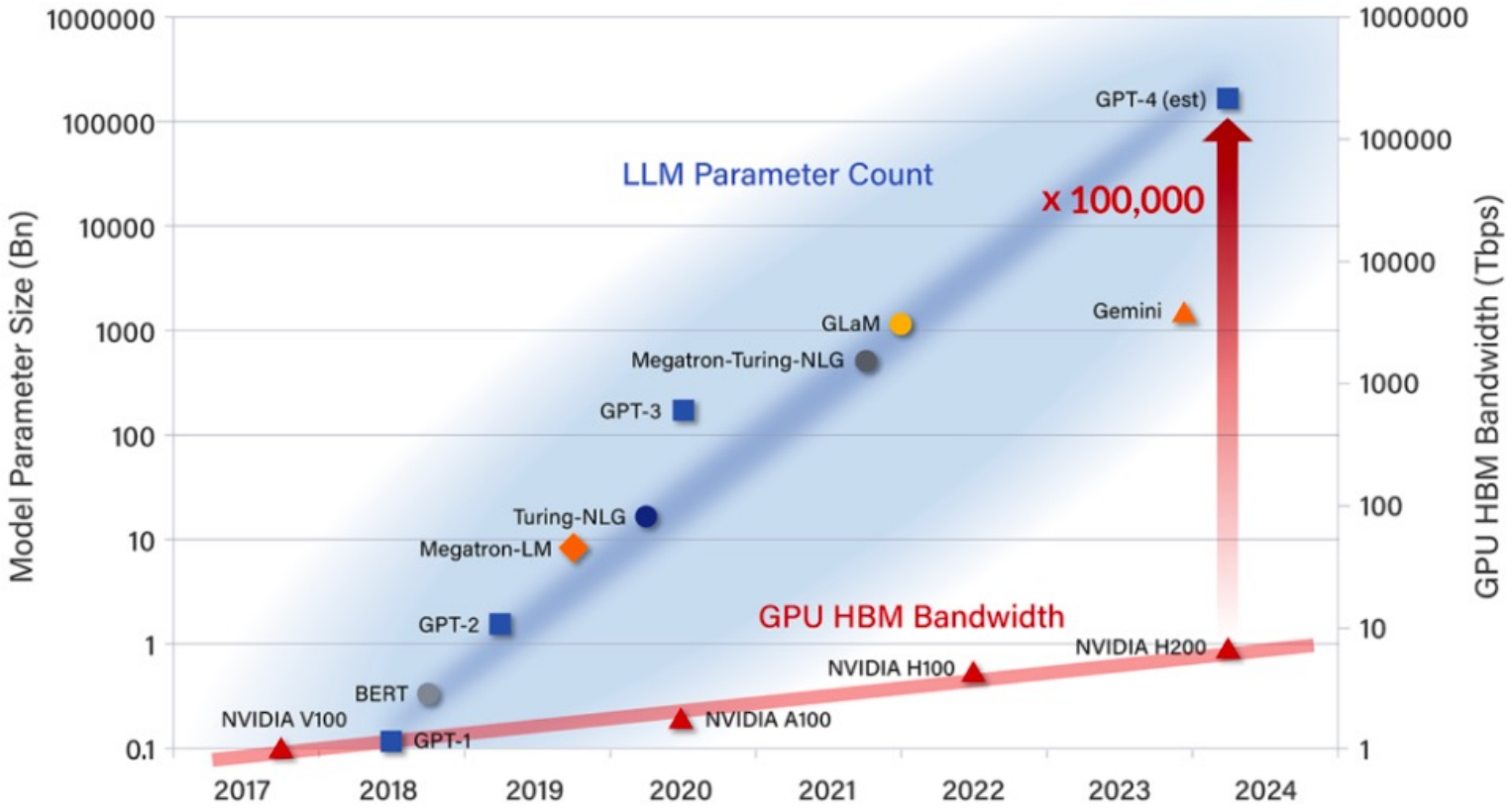


FP16, TF16, BF16

Introduction

Quantization LLM의 시대

GPT-3 : 1,750억 개, 하이퍼클로바 : 2,000억 개, LG AI연구원 엑사원 : 3,000억 개



크기가 증가할수록 도메인 특화 및 배포 등 서비스에 대한 부담이 커지고 높은 에너지 소비로 인한 환경 및 경제적 영향에 대한 우려가 커짐

→ 이를 해결하기 위해 양자화 사용

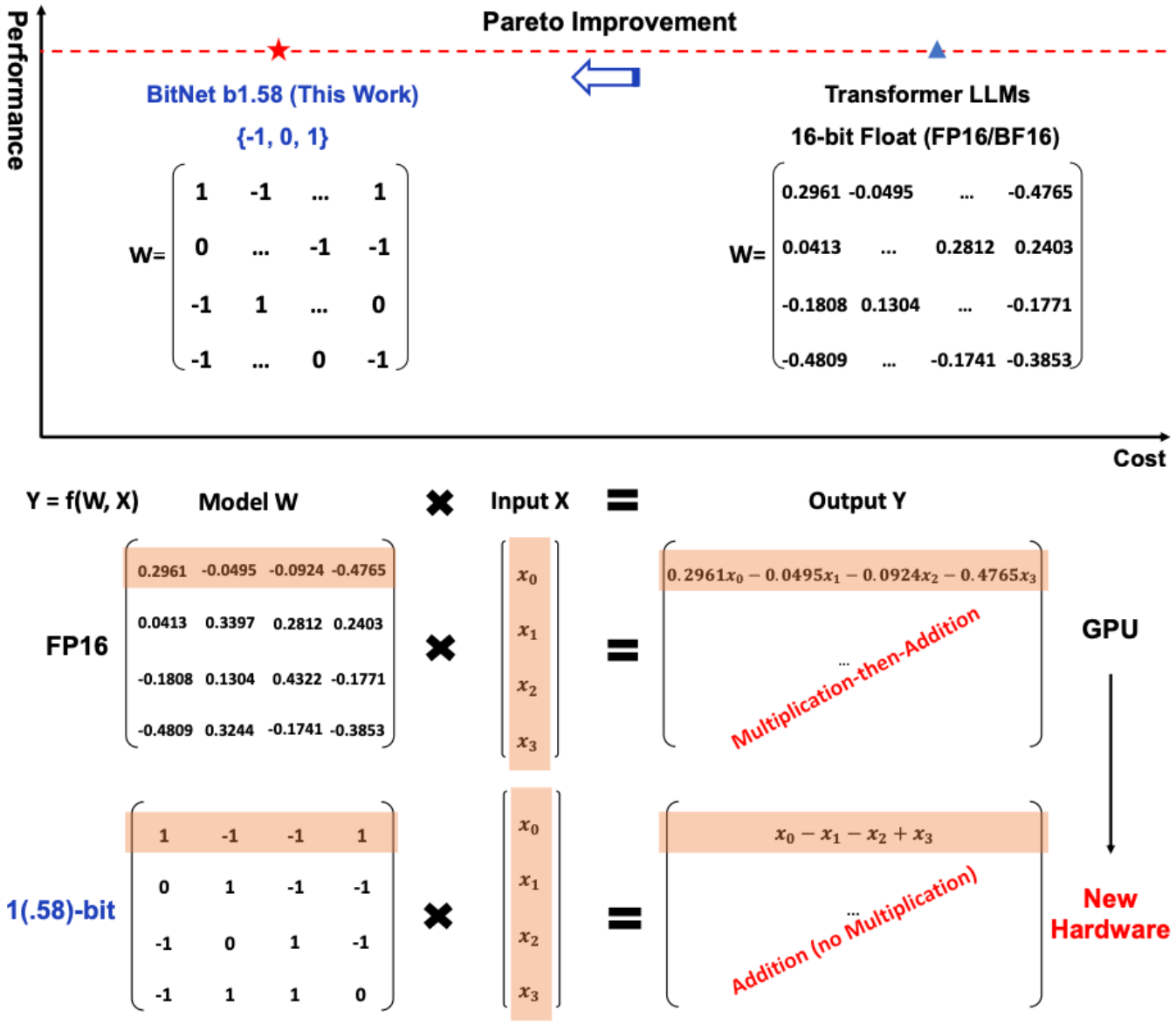
Stanford Alpaca

Introduction

현재는 16bit ⟶ 4bit quantization이 트렌드(LoRa, QLoRa)

- 1Bit LLM 변형인 BitNet b1.58 : **파라미터가 삼진법(-1, 0, 1)을 사용**하여 계산 비용을 줄이고 모델의 효율성을 향상시킨다.
- 비용 효율성 : BitNet b1.58은 기존 LLM에 비해 지연 시간, 메모리 사용량, 처리량, 에너지 소비 측면에서 상당한 비용 절감을 제공한다. 이는 **더 낮은 비용으로 동일하거나 더 나은 성능을 달성**할 수 있음을 의미한다.
- 새로운 스케일링 법칙과 훈련 방법 : 1.58 bit LLM은 **고성능임에도 효율적인 비용**의 새로운 세대의 LLM을 훈련하는 새로운 법칙과 방법을 제시한다.
- 새로운 계산 패러다임 : 거의 모든 행렬 곱셈 연산에서 **곱연산을 필요로 하지 않으므로**, 계산을 크게 단순화하고 최적화할 수 있는 새로운 계산 패러다임을 제공한다.
- 1Bit LLM을 위한 특정 하드웨어 설계 : BitNet b1.58의 계산은 1bit LLM에 최적화된 특정 하드웨어를 설계할 수 있는 새로운 가능성을 연다.

Introduction



- 파레토 개선
  - Performance는 유지하면서 Cost가 낮아짐
- 연산 방법
  - 기존 FP16 연산 시 부동소수점 곱셈
  - 1.58bit 계산 시 정수 덧셈만

파레토 개선(Pareto Improvement)
- 한 상태에서 다른 상태로 변화할 때 적어도 한 명은 상태가 개선되고 나머지는 악화되지 않는 경우

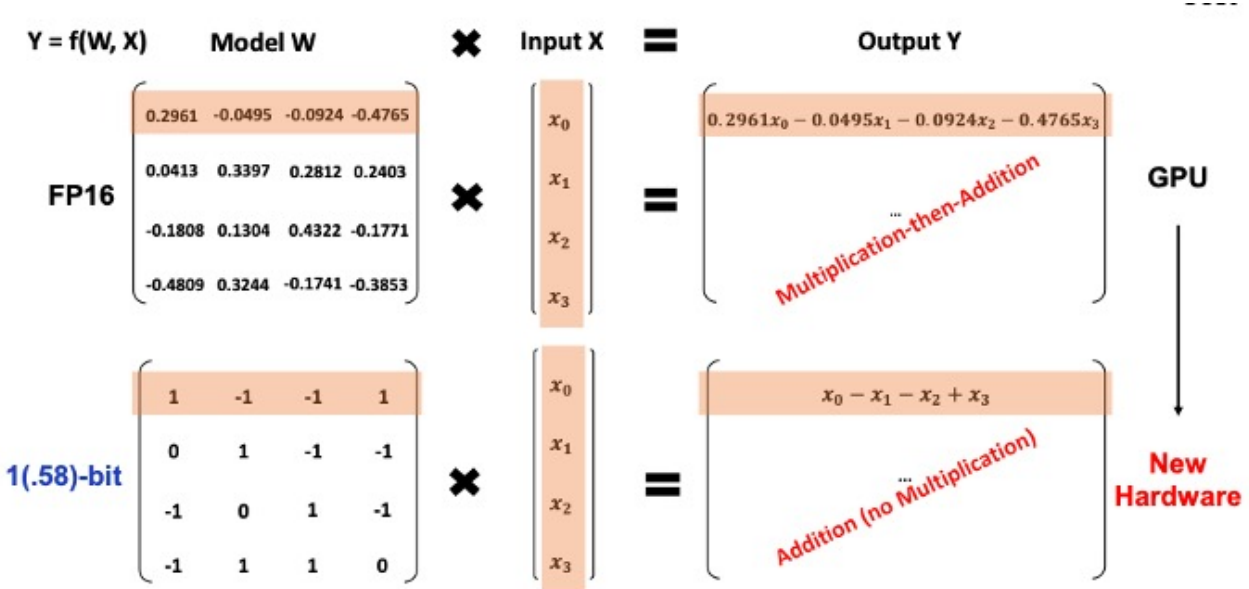
BitNet b1.58

1-bit BitNet vs 1.58-bit BitNet

nn.linear를 BitLinear로 대체하는 Transformer인 ButNet Architecture에 기반.

1.58bit 가중치와 8bit activation으로 훈련

1bit BitNet(가중치를 1-bit로 양자화)에 0을 추가하여 1.58bit로 변경. 기존 1bit의 이점을 유지하며, 0을 추가해 연산 성능을 향상시킴.



BitNet b1.58

Quantization Function

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

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

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

Absmean Quantization Function(절대 양자화 함수)을 적용해 가중치를 {-1, 0, 1}로 제한

1. weight matrix(가중치 행렬)를 그것의 average absolute value(평균 절대값)으로 scaling
2. 각 value를 {-1, 0, 1} 중 가장 가까운 값으로 반올림
3. 평균 절대값 계산 : γ는 가중치 행렬의 평균 절대값을 나타내며, 모든 가중치 값의 절대값의 합을 가충치의 총 수로 나눈 값

Results

Llama 채택
RMSNorm, SwiGLU, rotary embedding 사용,
모든 bias 제거

BitNet b1.58은 pretrain시키는 방법으로 가능하기 때문에 현재 우리가 사용할 수는 없음.

3B인 경우 2.71배 더 빠르고 3.55배 더 적은 메모리 사용.
특히 3.9B인 경우 2.4배 빠르고 메모리는 3.32배 더 적게 사용하지만 성능은 향상된걸 볼 수 있음

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**

Results

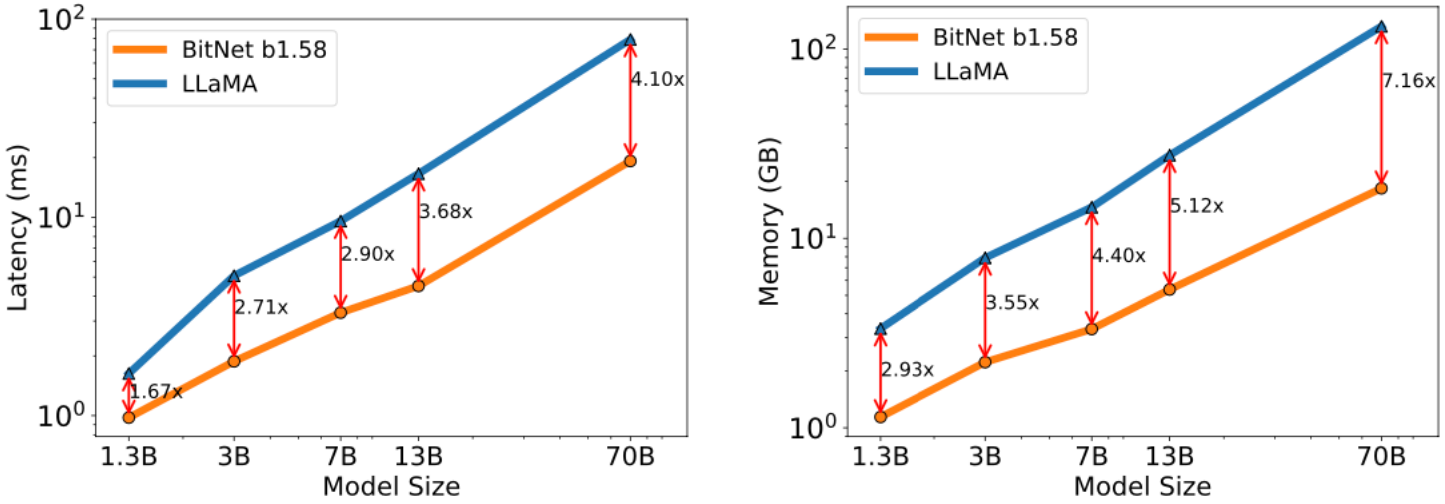


Figure 2: Decoding latency (Left) and memory consumption (Right) of BitNet b1.58 varying the model size.

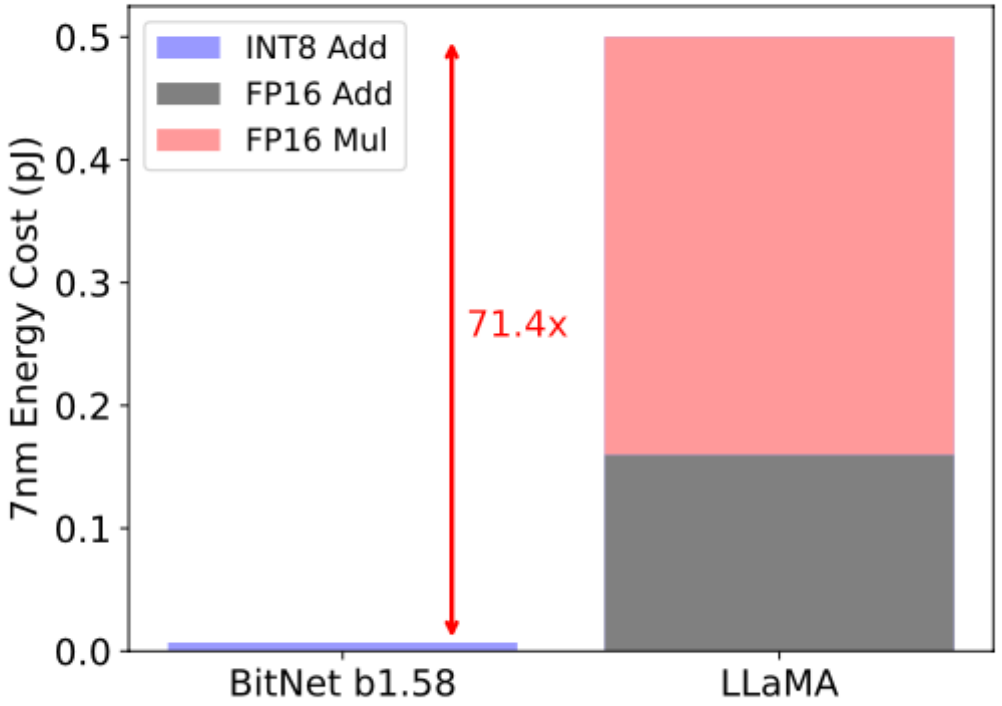
Decoding Latency와 memory consumption 모두 BitNet b1.58이 큰 차이를 보이는 것을 볼 수 있다.

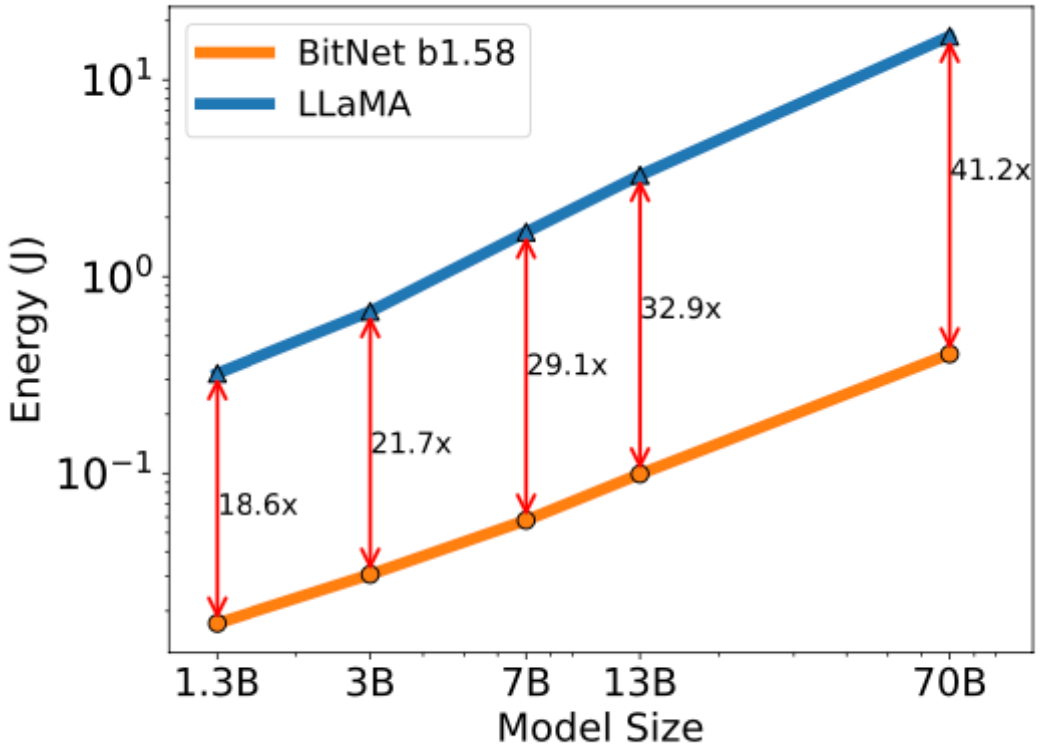
Models	Size	Max Batch Size	Throughput (tokens/s)
LLaMA LLM	70B	16 (1.0x)	333 (1.0x)
BitNet b1.58	70B	**176 (11.0x)**	**2977 (8.9x)**

Table 3: Comparison of the throughput between BitNet b1.58 70B and LLaMA LLM 70B.

Throughput(모델이 시간당 처리할 수 있는 작업의 양) 또한 BitNt b1.58이 더 좋은 성능을 보여주고 있다.

Results


INT8 Add
FP16 Add
FP16 Mul
7nm Energy Cost (pJ)
0.5
0.4
0.3
0.2
0.1
0.0
71.4x
BitNet b1.58
LLaMA


BitNet b1.58
LLaMA
Energy (J)
10^1
10^0
10^{-1}
18.6x
21.7x
29.1x
32.9x
41.2x
1.3B
3B
7B
13B
70B
Model Size

Discussion and Future Work

1-bit Mixture-of-Experts(MoE) LLMs

Mixture-of-Experts(MoE) : LLM의 cost를 효과적으로 다루는 방법

MoE는 FLOPs(초당 부동 소수점 연산의 수)를 줄이는 대신 high memory consumption과 inter-chip communication 때문에 사용화하기는 어렵다.

하지만 b1.58로 해결할 수 있다.

- 1. 줄어든 memory footprint가 MoE 모델 배포를 더 쉽게 만들어준다.
- 2. 네트워크간 활성화 전송이 overhead(연산의 추가적인 부담)를 크게 줄여준다.
- 3. 전체 모델을 단일칩에 배치하여 overhead 자체를 없앤다.

Native Support of Long Sequence in LLMs

Long Sequence를 처리하는 것은 LLM의 주요과제 중 하나다.
- Sequence가 길수록 memory 소모가 크기 때문.

1.58 bit를 사용하면 이를 해결할 수 있다.

Discussion and Future Work

LLMs on Edge and Mobile

edge와 mobile device에서의 LLM 사용에 크게 기여한다.(작은 장치에서 고성능 LLM 사용 가능)

New Hardware for 1-bit

LLM 특화 Hardware가 개발되었는데 1-bit Model에 특화된 Hardware도 기대할 수 있다.

올해 Grop에서 보여준LPU는 LLM을 위해 설계된 하드웨어의 성능을 잘 보여주었다. 이와 통합해 edge or mobile에서도 고성능 LLM을 사용할 수 있게 될 것으로 기대한다.

Conclusion

1. 기존 LLM이 가지는 문제(고에너지, 고메모리 필요 등)를 해결하는 BitNet b1.58 제안
2. 파라미터를 삼진법{-1, 0, 1}으로 사용하여 연산 성능 향상
3. sLLM 중 하나인 Llama와 비교했을 때 좋은 성능을 보여줌