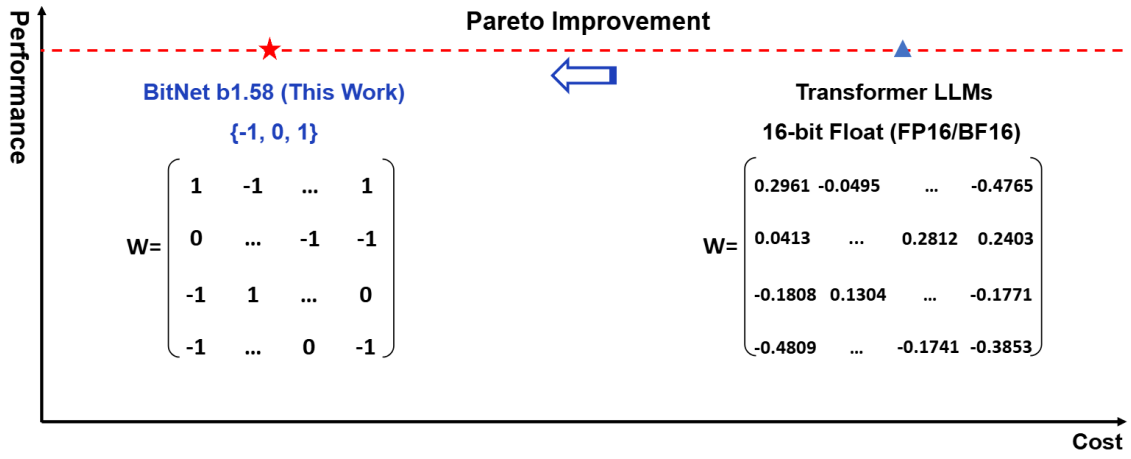
The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

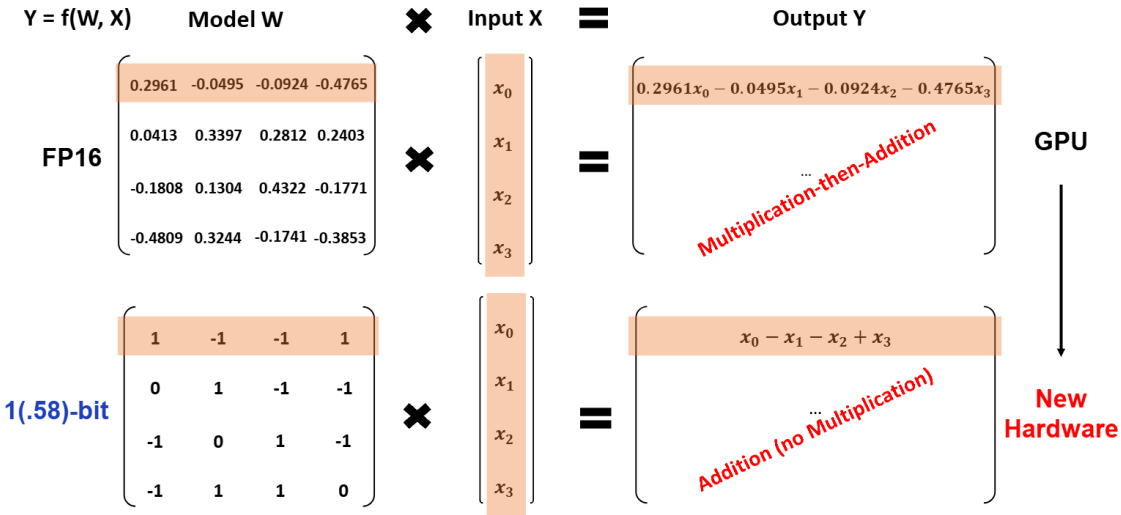
길종현(github.com/hyeon-n-off)

개요

LLM 의 모든 파라미터들이 ternary $\{-1,0,1\}$ 로 이루어진 1-bit LLM 의 변형 **BitNet b1.58** 을 소개한다. 전체 정밀도(full-precision, i.e. FP16 or BF16) Transformer 구조의 LLM 들과 같은 모델 크기와 훈련 토큰을 사용하여 지연 시간, 메모리, 처리량, 에너지 소비량 면에서 훨씬 효율적인 결과를 나타내었다.

더 심오하게, 1.58-bit LLM 은 고성능이면서 비용 효율적인 새로운 스케일링 법칙(scaling law)과 차세대 LLM 의 훈련 방법을 정의한다. 또한 새로운 계산 패러다임을 구현하고 1-bit LLM 에 최적화된 특정 하드웨어를 설계할 수 있는 기회를 열어준다.





The Era of 1-bit LLMs

최근 몇 년동안, AI 분야에서 LLMs 의 규모와 성능이 급성장을 하였다. 이러한 모델들은 광범위한 NLP task 에서 훌륭한 성능을 보여주었지만, 모델의 증가한 크기는 배포를 어렵게 하고, 높은 에너지 소비로 인해 환경 및 경제적 영향에 대한 우려를 높아지게 하였다. 이러한 문제를 해결하기 위한 방법은 PTQ(Post Training Quantization)을 사용하여 추론을 위한 low-bit 모델들을 생성하는 것이다. 이는 LLMs 의 가중치와 활성화값의 정밀도를 낮춰 메모리와 계산 요구량을 효과적으로 줄일 수 있다. 그러나 PTQ 는 LLMs 에 널리 쓰이긴 하지만, 차선책(sub-optimal)에 불과하다.

BitNet 과 같은 최근 1-bit 모델의 연구는 성능을 유지하면서 LLMs 의 비용을 줄일 수 있는 방향을 제시한다. 기존의 LLMs 은 16-bit 부동 소수점을 사용하였고, LLMs 대부분은 행렬 곱으로 이루어져 있다. 따라서 대부분의 계산 연산량은 부동 소수점 덧셈과 곱셈 연산에서 온다. 대조적으로, BitNet 의 행렬 곱은 오직 정수 덧셈(integer addition)만을 포함한다. 많은 칩들에서 컴퓨팅 성능의 근본적인 한계는 전력이므로 에너지 절약은 더 빠른 컴퓨팅으로 이어질 수 있다.

계산 외에도, 모델 파라미터를 DRAM 에서 on-chip accelerator 의 메모리(e.g. SRAM)로 전송하는 과정은 추론 중에 비용이 많이 들 수 있다. 처리량을 향상시키기 위해 SRAM 의 크기를 확대려하는 시도가 있었지만 이는 DRAM 보다 훨씬 더 높은 비용을 초래한다. 완전 정밀도 모델에 비해 1-bit LLMs 는 용량 및 대역폭 측면에서 메모리 공간이 훨씬 적다. 이를 통해 DRAM 에서 가중치를 로드하는데 드는 비용과 시간을 크게 줄여 더 빠르고 효율적인 추론을 가능하게 한다.

이 논문에서, 모든 파라미터가 tenary 인 1-bit LLM 의 변형 **BitNet b1.58** 을 소개한다. 원래 1-bit BitNet 에서 '0' 값을 추가하여 이진 시스템에서 1.58bit 가 되었다. BitNet b1.58 은 BitNet 의 모든 장점을 가지며, 행렬 곱셈을 위한 연산이 거의 필요하지 않으며 고도로 최적화될 수 있는 새로운 계산 패러다임을 포함한다.

BitNet b1.58

BitNet b1.58 은 *nn.Linear* 를 *BitLinear* 로 교체한 Transformer 구조의 BitNet 구조에 기반하였다.

1.58-bit 가중치와 8-bit 활성화 값을 사용하여 훈련되었다. 원래 BitNet 과 비교하여 아래에 요약된 몇 가지 수정 사항이 도입되었다.

Quantization Function

가중치를 -1, 0, 1 로 제한하기 위해 ***절대값 양자화(absmean quantization)*** 함수를 채택한다.

먼저 평균 절댓값으로 가중치 행렬의 크기를 조정(scale)한 다음, 각 값을 {-1, 0, 1} 중에서 가장 가까운 정수로 반올림한다.

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

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

$$\gamma = \frac{1}{nm}\sum_{ij}|W_{ij}| \qquad (3)$$

활성화에 대한 양자화 함수는 비선형 함수 이전의 활성화를 $[0, Q_b]$ 범위로 조정(scale)하지 않는다는 점을 제외하면 BitNet 과 동일한 구현을 따른다. 대신 활성화는 zero-point 양자화를 제거하기 위해 토큰 당 모두 $[-Q_b, Q_b]$ 범위로 조정된다. 이는 구현과 시스템 수준의 최적화 모두에 대해 보다 편리하고 간단하다.

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.

LLaMA-alike Components

LLaMA 의 아키텍처는 오픈 소스 LLM 의 사실상 중추 역할을 해왔다. 오픈 소스 커뮤니티를 수용하기 위해 BitNet b1.58 의 설계는 LLaMA 와 유사한 구성 요소를 채택하고 있다. 특히, RMSNorm, SwiGLU, rotary embedding 을 사용하여 모든 변향을 제거한다.

Results

BitNet b1.58 을 다양한 크기로 재현된 FP16 LLaMA LLM 와 비교하였다. 공정한 비교를 위해 1,000 억 개의 토큰에 대해 RedPajama 데이터셋에서 사전 훈련을 진행하였다. 런타임 GPU 메모리와 지연 시간을 비교하였다.

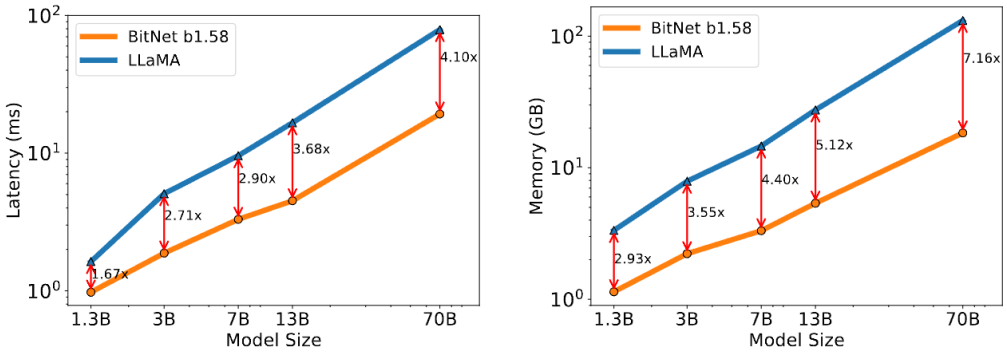


Figure 2: Decoding latency (Left) and memory consumption (Right) of BitNet b1.58 varying the model size.

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.

표 2 의 결과는 모델의 크기가 증가함에 따라 BitNet b1.58 과 LLaMA LLM 간의 성능 격차가 좁아지는 것을 보여준다. 더 중요한 것은 BitNet b1.58 이 전체 정밀도 baseline 의 성능과 일치할 수 있다는 것이다. 최종 작업 결과는 BitNet b1.58 3.9B 가 더 낮은 메모리 및 지연 시간 비용으로 LLaMA LLM 3B 보다 성능이 우수하다는 것을 보여준다.

Memory and Latency 모델 크기를 7B, 13B, 70B 로 더욱 확장하여 평가하였다. 그림 2 에서 지연 시간과 메모리의 경향을 보여주며, 모델 크기가 확장됨에 따라 속도 향상이 증가함을 보여준다. 이는 *nn.Linear* 의 시간 비용이 모델 크기에 따라 증가하기 때문이다. 대기 시간과 메모리 모두 2-bit 커널로 측정되었으므로 비용을 더욱 절감하기 위한 최적화의 여지는 아직 남아있다.

Energy LLM 비용에 가장 큰 영향을 주는 행렬 곱셈 계산에 주로 중점을 두었다. 그림 3 은 에너지 비용의 구성을 보여준다. BitNet b1.58 의 대부분은 INT 덧셈 계산인 반면, LLaMA LLM 은 FP16 덧셈과 FP16 곱셈으로 구성된다. [Hor14, ZZL22]의 에너지 모델에 따르면 BitNet b1.58 은 7nm 칩에서 행렬 곱셈을 위한 산술 연산 에너지 소비를 약 71.4 배 절약한다. 모델 크기가 확장됨에 따라 에너지 소비 측면에서 점점 더 효율적이게 된다는 것을 보여준다.

Throughput 시퀀스 길이가 512 인 GPU 메모리 제한에 도달할 때까지 배치 크기를 늘렸다. 표 3 은 BitNet b1.58 70B 가 LLaMA LLM 의 배치 크기를 최대 11 배까지 지원하여 처리량을 8.9 배 더 높일 수 있음을 보여준다.

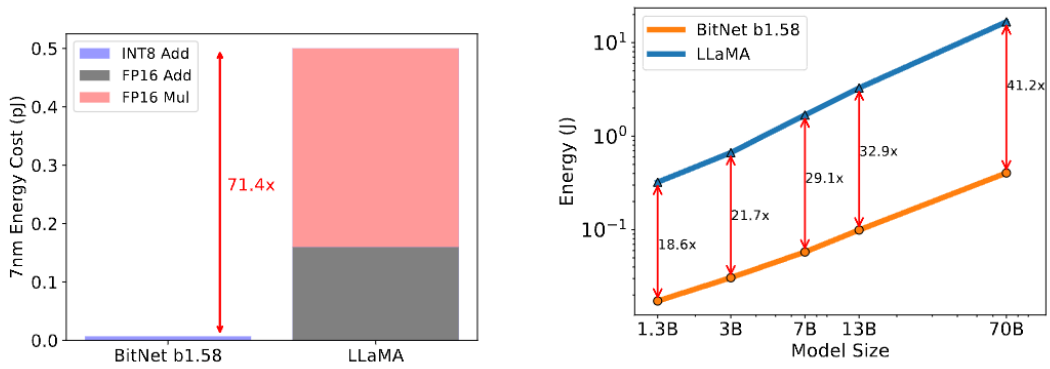


Figure 3: Energy consumption of BitNet b1.58 compared to LLaMA LLM at 7nm process nodes. On the left is the components of arithmetic operations energy. On the right is the end-to-end energy cost across different model sizes.

Models	Tokens	Winogrande	PIQA	SciQ	LAMBADA	ARC-easy	Avg.
StableLM-3B	2T	64.56	76.93	90.75	66.09	67.78	73.22
BitNet b1.58 3B	2T	**66.37**	**78.40**	**91.20**	**67.63**	**68.12**	**74.34**

Table 4: Comparison of BitNet b1.58 with StableLM-3B with 2T tokens.