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

2024

Shuming Ma*, Hongyu Wang*, Lingxiao Ma, Lei Wang, Wenhui Wang,
Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, Furu Wei⋄

∗ Equal contribution. ⋄ Corresponding author. S. Ma, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, J. Xue, F.Wei are with Microsoft Research. H.Wang and R.Wang are with University of Chinese Academy of Sciences.

Part 1.

Introduction

- **Quantized Network in Deep Learning**
    - Vanilla LLMs
        - 16-bit floating values (i.e., FP16 or BF16)
        - The bulk of any LLMs is matrix multiplication
    - BitNet [WMD+23]
        - Matrix multiplication of BitNet only involves integer addition
        - Reduce the cost and time of loading weights from DRAM
        - More energy saving can be translated into faster computation and more efficient inference
    - BitNet b1.58
        - 1-bit LLM, where every parameter is ternary, taking on values of {-1, 0, 1}
        - Additional value of 0 to {-1, 1} results in 1.58 bits in the binary system
        - Require almost no multiplication operations for matrix multiplication and can be highly optimized

Hongyu Wang et al. Bitnet: Scaling 1-bit transformers for large language models. CoRR, abs/2310.11453, 2023.

Part 2.

Experiment

- **Implementation Details**
  - Model
    - BitNet b1.58 to our reproduced FP16 LLaMA LLM in various sizes
    - Pre-trained the models on the RedPajama dataset
  - Dataset
    - Zero-shot performance on ARC-Easy, ARC-Challenge, Hellaswag, Winogrande, PIQA, OpenbookQA, BoolQ
    - Validation perplexity on WikiText2, C4
  - Comparison of the runtime GPU memory and latency
    - Results were measured using the FasterTransformer codebase
    - 2-bit kernel from Ladder [WMC+23] is also integrated for BitNet b1.58
    - Report the time per output token, as it is the major cost for inference.

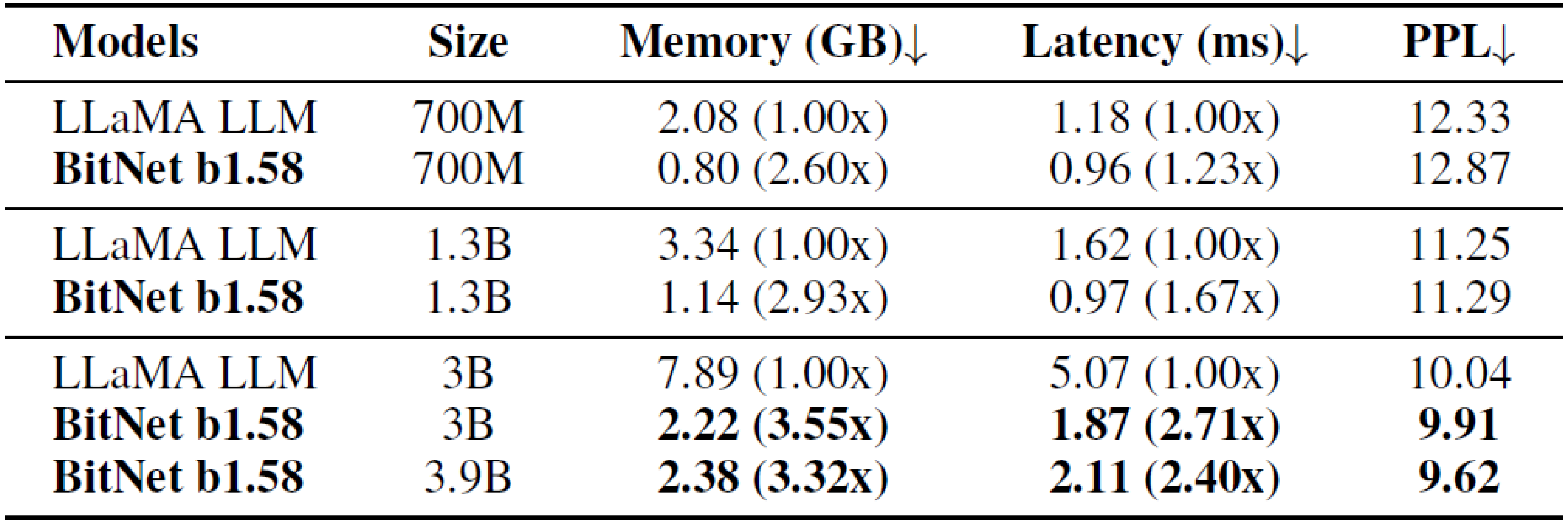
https://github.com/NVIDIA/FasterTransformer

LeiWang et al. Ladder: Efficient tensor compilation on customized data format. OSDI. 2023.

Part 2.

Experiment

- **Perplexity & Cost**
  - 3B model size
    - BitNet b1.58 starts to match full precision LLaMA LLM in terms of perplexity
    - 2.71 times faster and 3.55 times less GPU memory
  - 3.9B model size
    - Performs significantly better than LLaMA LLM 3B
    - 2.4 times faster, 3.32 times less GPU memory

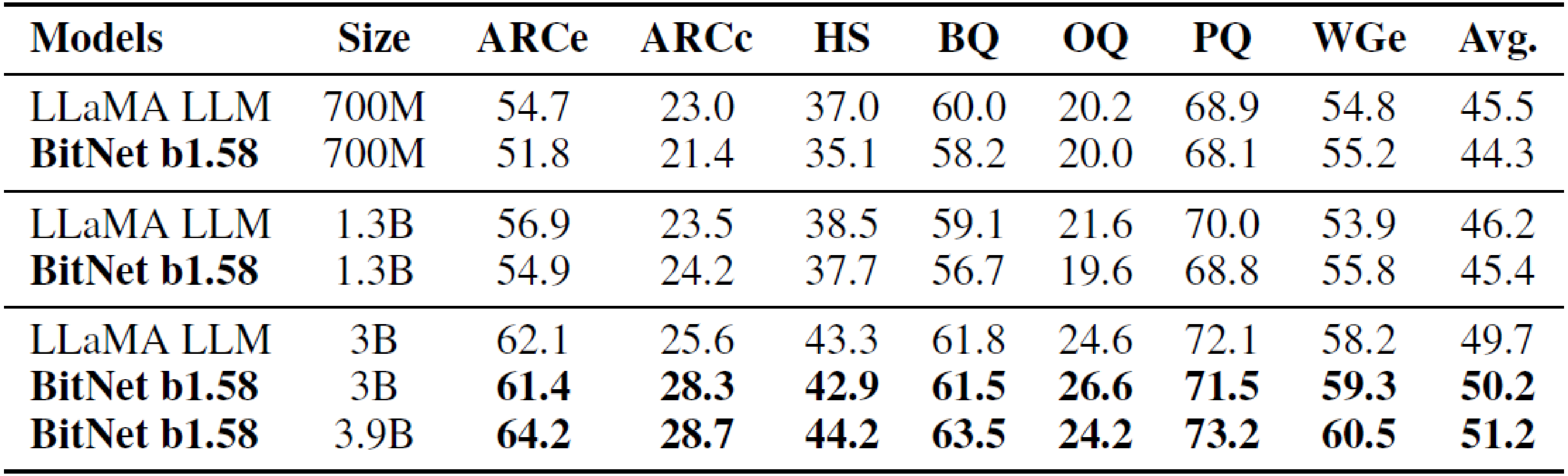


Models	**Size**	**Memory (GB)↓**	**Latency (ms)↓**	**PPL↓**
LLaMA LLM	700M	2.08 (1.00x)	1.18 (1.00x)	12.33
BitNet b1.58	700M	0.80 (2.60x)	0.96 (1.23x)	12.87
LLaMA LLM	1.3B	3.34 (1.00x)	1.62 (1.00x)	11.25
BitNet b1.58	1.3B	1.14 (2.93x)	0.97 (1.67x)	11.29
LLaMA LLM	3B	7.89 (1.00x)	5.07 (1.00x)	10.04
BitNet 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**

Part 2.

Experiment

- **Zero-shot accuracy on the end tasks Memory and Latency**
  - Follow the pipeline from lm-evaluation-harness4 to perform the evaluation
  - Performance gap between BitNet b1.58 and LLaMA LLM narrows as the model size increases
  - Similar to the observation of the perplexity, BitNet b1.58 3.9B outperforms LLaMA LLM 3B with lower memory and latency cost



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**

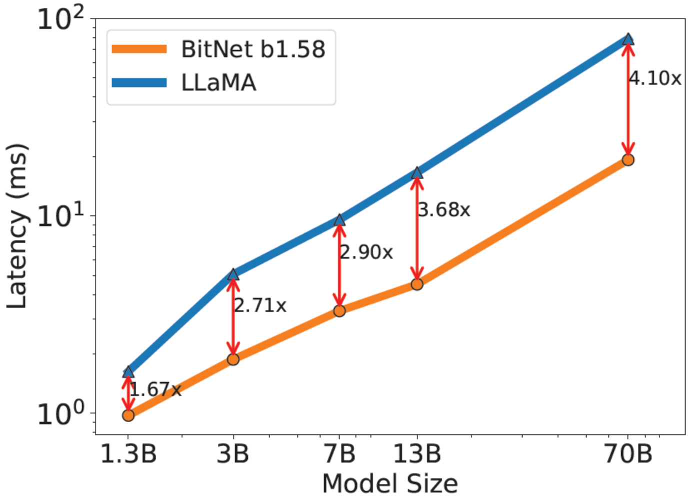
https://github.com/EleutherAI/lm-evaluation-harness

Experiment

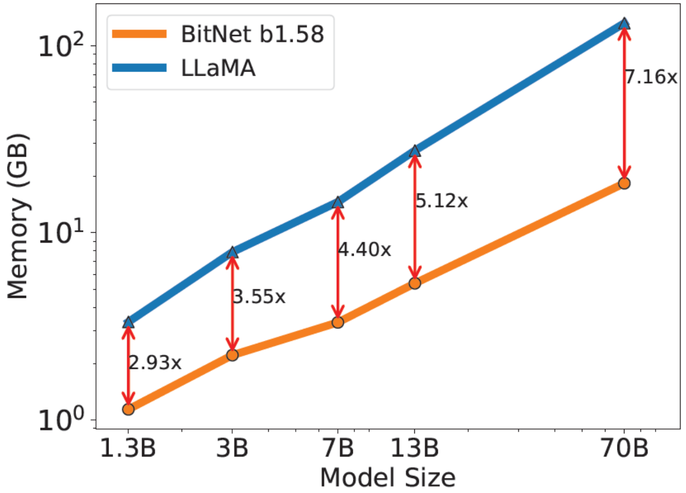
• Latency & Memory

- Latency
  - BitNet b1.58 70B is 4.1 times faster than the LLaMA LLM baseline
  - This is because time cost for nn.Linear grows with the model size
- Memory
  - Embedding remains full precision and its memory proportion is smaller for larger models
  - Measurement unit is a 2-bit kernel

Decoding latency



Memory consumption



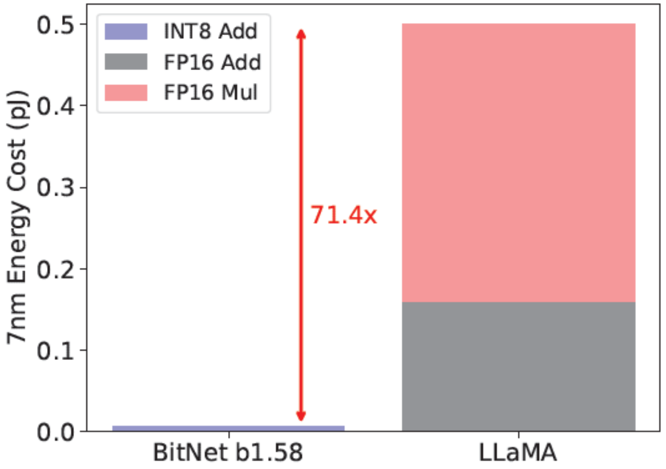
Hongyu Wang et al. Bitnet: Scaling 1-bit transformers for large language models. CoRR, abs/2310.11453, 2023.

Experiment

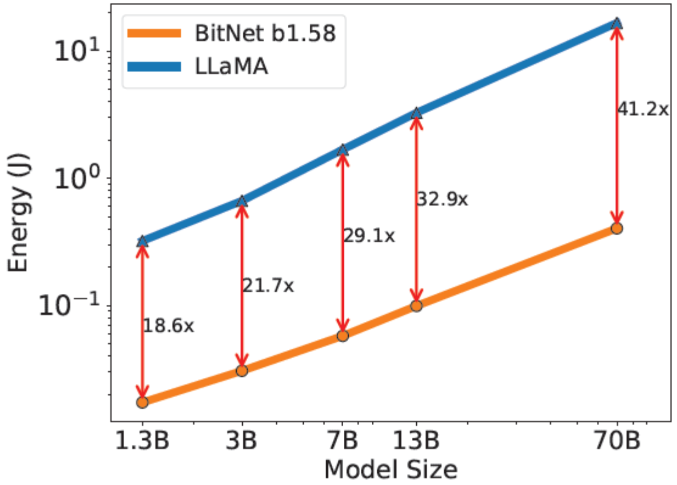
• Arithmetic operations energy consumption

- Focus mainly on the calculation for matrix multiplication
  - BitNet b1.58 saves 71.4 times energy consumption on 7nm chips
- As the model size scales, BitNet b1.58 becomes increasingly more efficient in terms of energy consumption
  - The percentage of nn.Linear grows with the model size
  - The cost from other components is smaller for larger models

Components of arithmetic operations energy



End-to-end energy cost

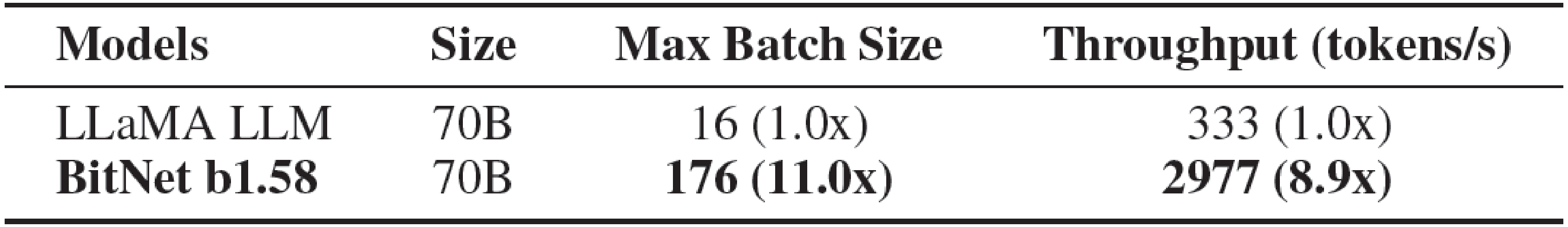


Part 2.

Experiment

- **Throughput**
  - Throughput of BitNet b1.58 and LLaMA LLM with 70B parameters on two 80GB A100 cards
  - Increase batch size until the GPU memory limit was reached, with a sequence length of 512
  - BitNet b1.58 70B can support up to 11 times the batch size of LLaMA LLM, resulting an 8.9 times higher throughput
- **Training with 2T Tokens**
  - Follow the data recipe of StableLM-3B [TBMR], which is the state-of-the-art open-source 3B model

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



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)**

Yanping Huang et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. NeurlPS. 2019.
Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. Stablelm 3b 4e1t.