The Era of 1-bit LLM

Speaker: Hongyu Wang

E-mail: why0711@mail.ustc.edu.cn

BitNet: Scaling 1-bit Transformers for Large Language Models

Hongyu Wang*[12], Shuming Ma*[1], Li Dong[1] , Shaohan Huang[1],
Huaijie Wang[3], LingXiao Ma[1], Fan Yang[1], Ruiping Wang[2], Yi Wu[3], Furu Wei[1]

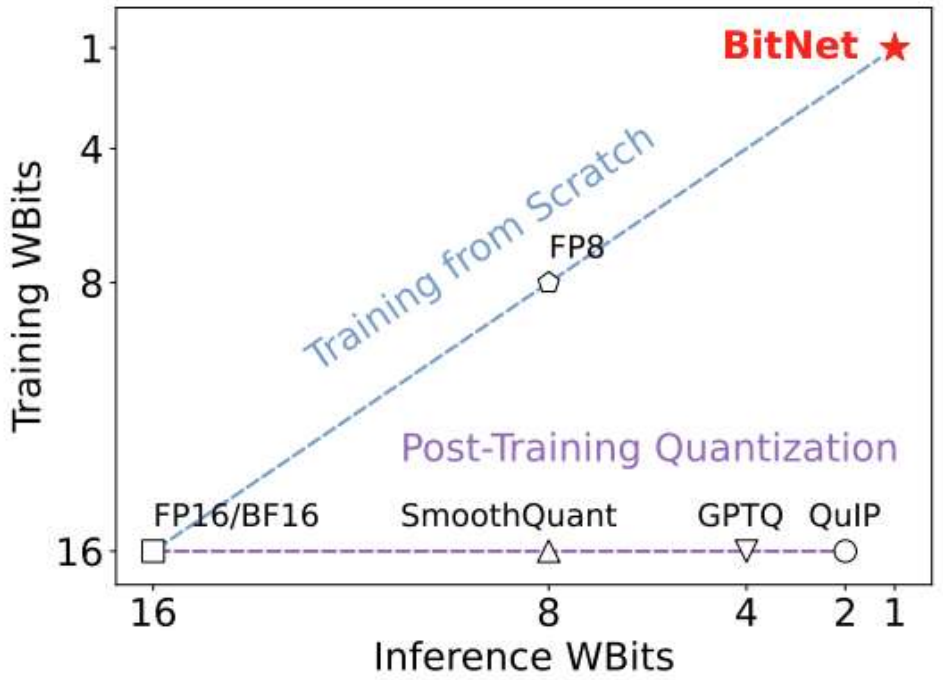
[1] Microsoft Research, [2] University of Chinese Academy of Sciences, [3] Tsinghua University

https://aka.ms/GeneralAI

Paper

01 Introduction

- The inference of LLM is memory-bound
- LLM is extremely sparse
  - Pruning
  - MoE
  - **Quantization**
- Quantization
  - Post-training Quantization
    - Pros: Small training cost.
    - Cons: Not work well on ultra-low precision(<2-bit) models
  - **Quantization-aware Training**



02 nn.Linear -> BitLinear

- Normalization*: Stabilize training (Sub-LN)

- Weight Quantization

$$\widetilde{W} = \mathrm{Sign}(W - \alpha),$$

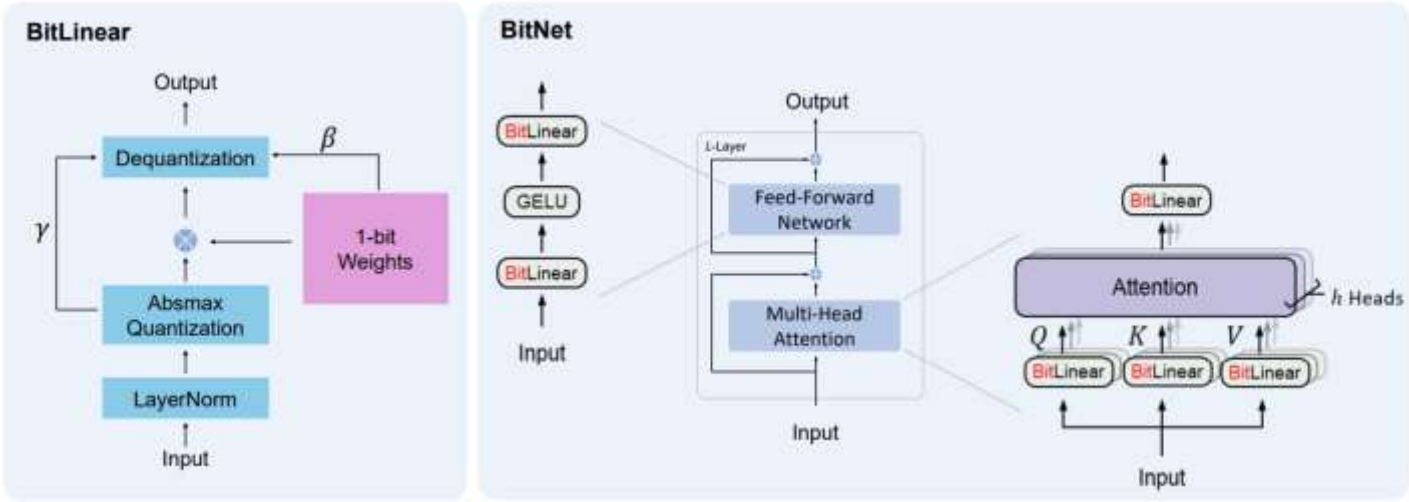
$$\alpha = \frac{1}{nm}\sum_{ij} W_{ij} \quad \beta = \frac{1}{nm}||W||_1$$

- Activation Quantization

$$\widetilde{x} = \mathrm{Quant}(x) = \mathrm{RoundClip}(\frac{Q_p}{\gamma + \epsilon}x, Q_n, Q_p)$$

$$\mathrm{RoundClip}(x, a, b) = \max(a, \min(b, \mathrm{round}(x)))$$

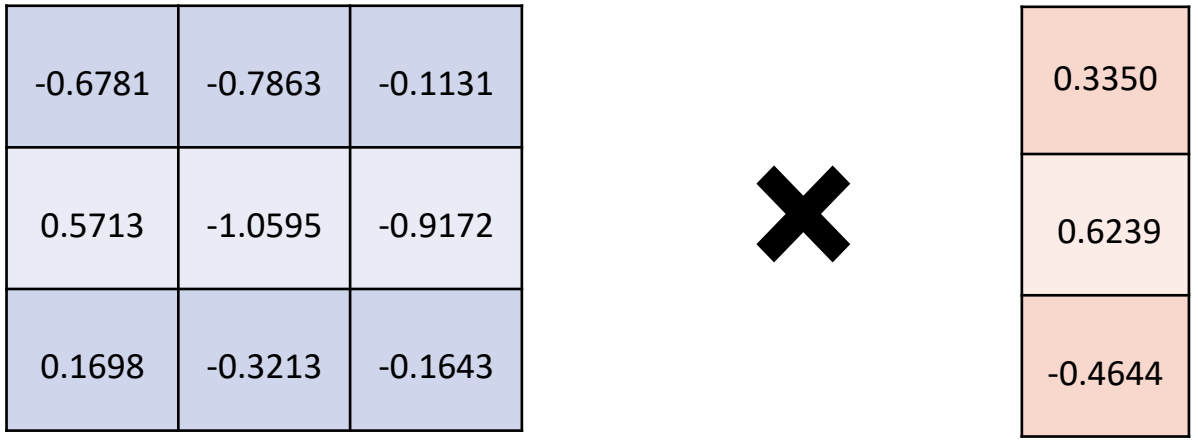
where $\gamma = ||x||_\infty$.



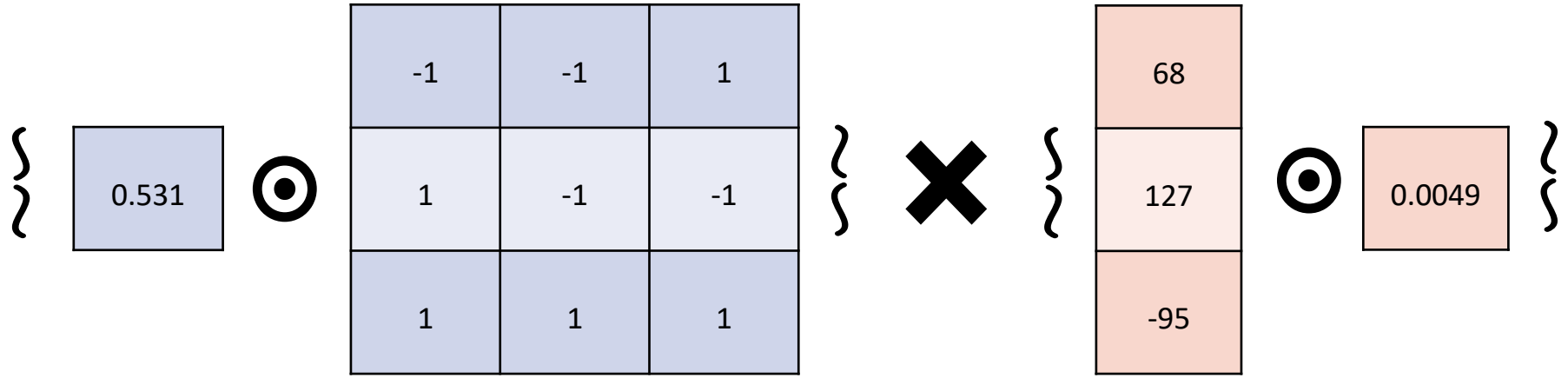
*Magneto: A Foundation Transformer. Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song and Furu Wei. ICML 2023

02 nn.Linear -> BitLinear

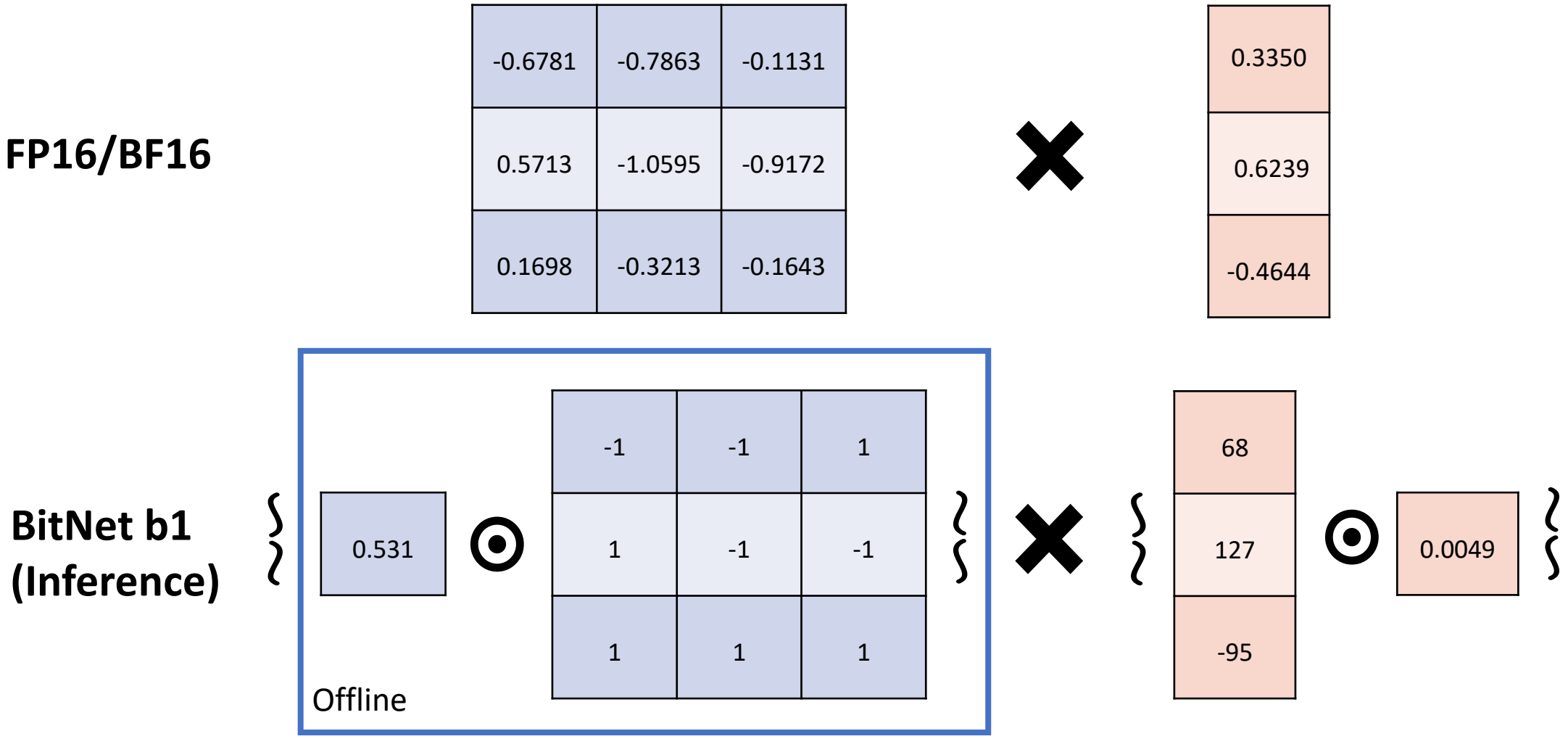
FP16/BF16



BitNet b1 (Training)



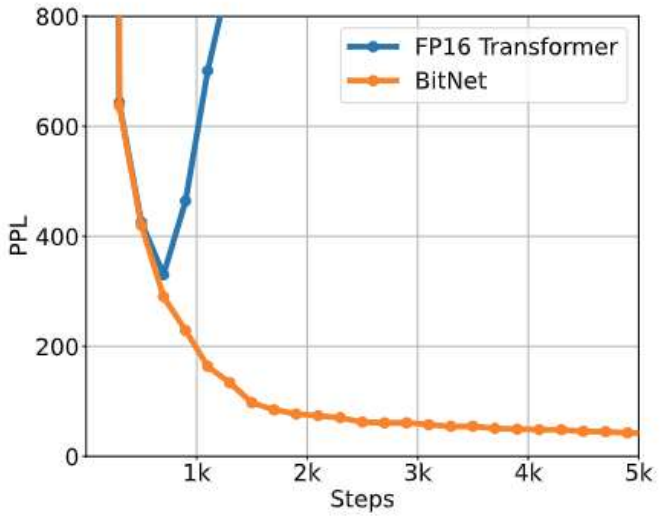
02 nn.Linear -> BitLinear



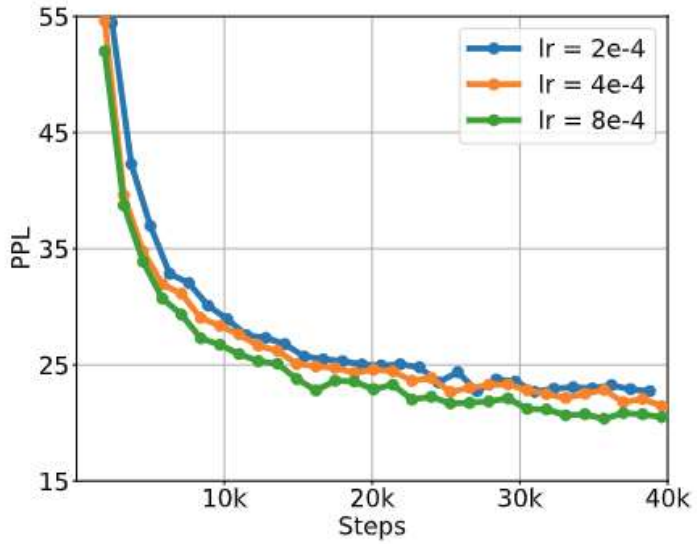
03 Training

- Gradient Estimation
  - Straight-through estimator(STE): Directly bypass the gradient of the weight and activations.
  - Maintain a latent weight (FP16) to accumulate parameter updates.

- Hyper-parameters
  - Large learning rate



BitNet is more stable than FP16 LLM



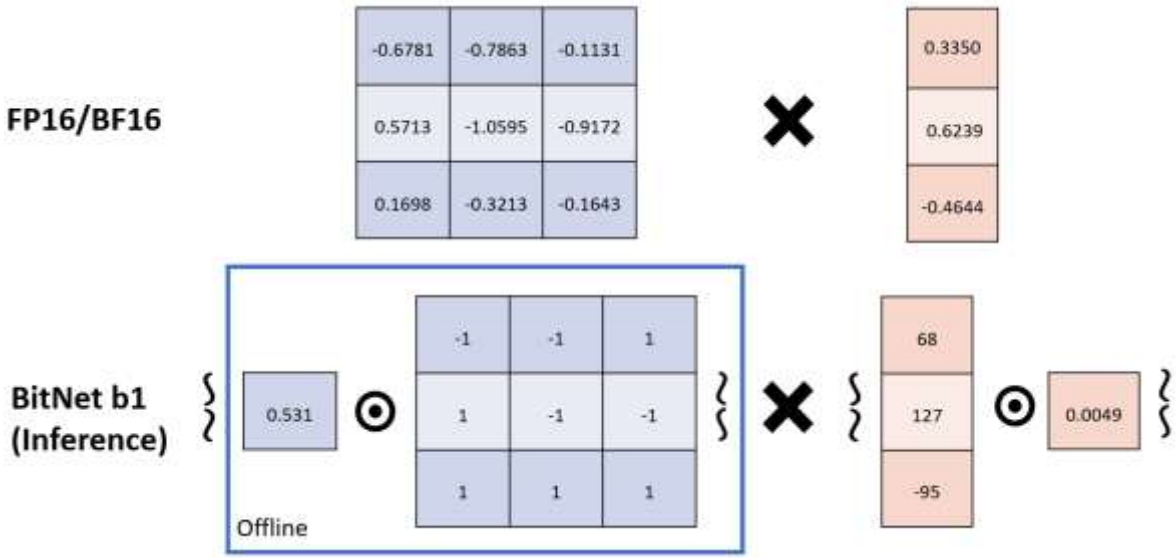
BitNet achieves lower PPL with larger learning rate.

04 Energy Consumption

- MatMul operation WX, $\mathrm{W} \in R^{\mathrm{n}\times n}, \mathrm{X} \in R^{\mathrm{n}\times 1}$

#Ops	FP16 ADD	FP16 MUL	INT8 ADD	INT8 MUL
FP16/BF16	$n(n-1)$	n^2	0	0
BitNet	0	$n+1$	$n(n-1)$	0

#Energy PerOp (pJ)	FP16 ADD	FP16 MUL	INT8 ADD	INT8 MUL
45nm	0.4	1.1	0.03	0.2
7nm	0.16	0.34	0.007	0.07

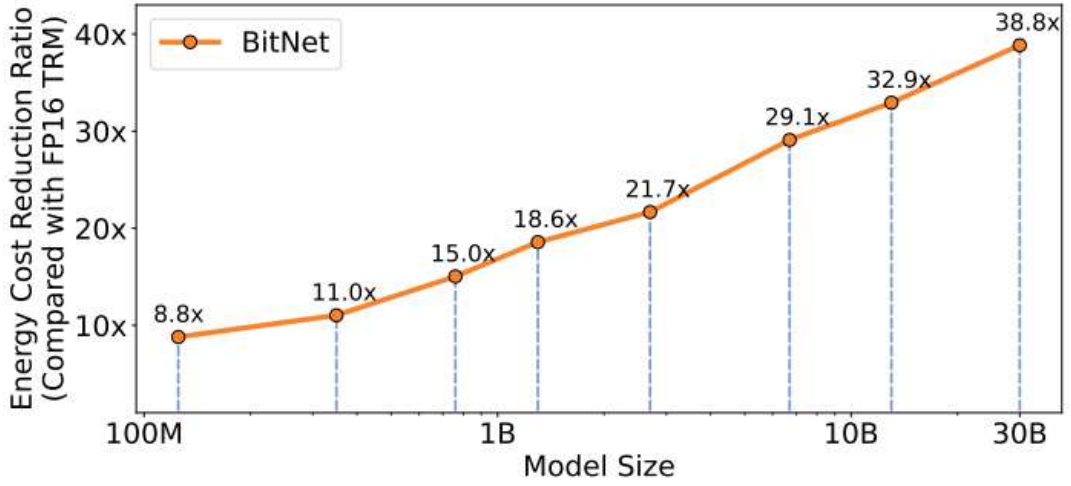


04 Energy Consumption

- MatMul operation WX, $\mathrm{W} \in R^{\mathrm{n}\times n}, \mathrm{X} \in R^{\mathrm{n}\times 1}$

#Ops	FP16 ADD	FP16 MUL	INT8 ADD	INT8 MUL
FP16/BF16	$n(n-1)$	n^2	0	0
BitNet	0	$n+1$	$n(n-1)$	0

- Energy cost ratio $\frac{\mathrm{FP16}}{\mathrm{BitNet}}$ increases as model size grows.

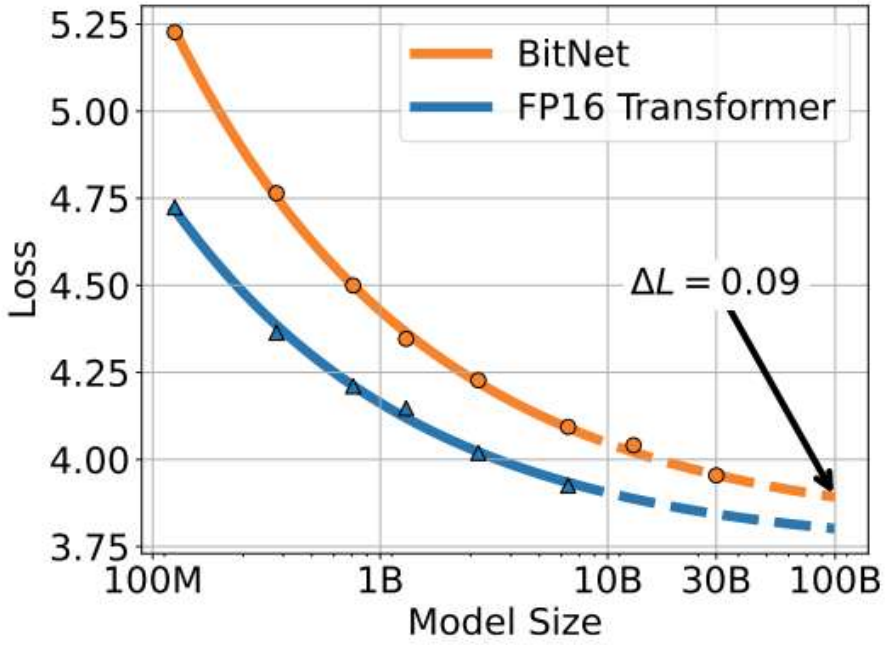


05 Scaling Law

- Does 1-bit LLM exhibit a scaling law? Yes!
  - The validation loss decreases as the model size grows.

- Is 1-bit LLM predictable? Yes!
  - We use the loss of 125M-2.7B model to predict the loss of 13B and 30B model.

- As the model size grows，the gap between 1-bit and FP16 LLM narrows.
  - 125M: $\Delta L = L_{fp16} - L_{1-bit} = 1.5$
  - 100B: $\Delta L = L_{fp16} - L_{1-bit} = 0.09$

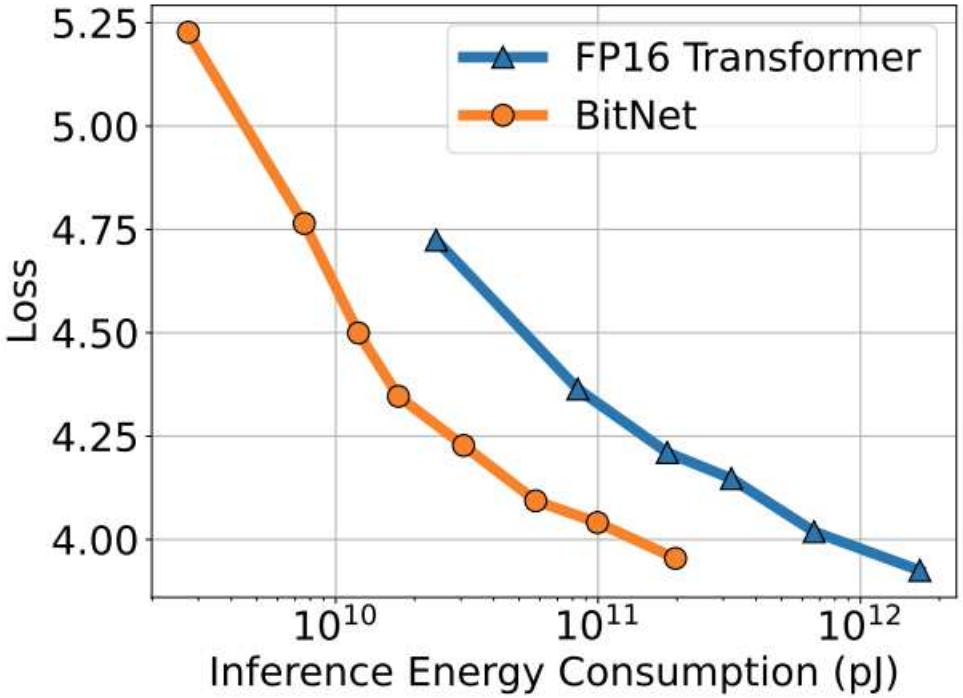


06 Inference-Optimal Scaling law

- Expected energy consumption

- BitNet achieves lower loss than FP16 LLM with the same inference energy budget.

- Energy: 30B BitNet ≈ 760M FP16 LLM



07 BitNet vs SoTA PTQs

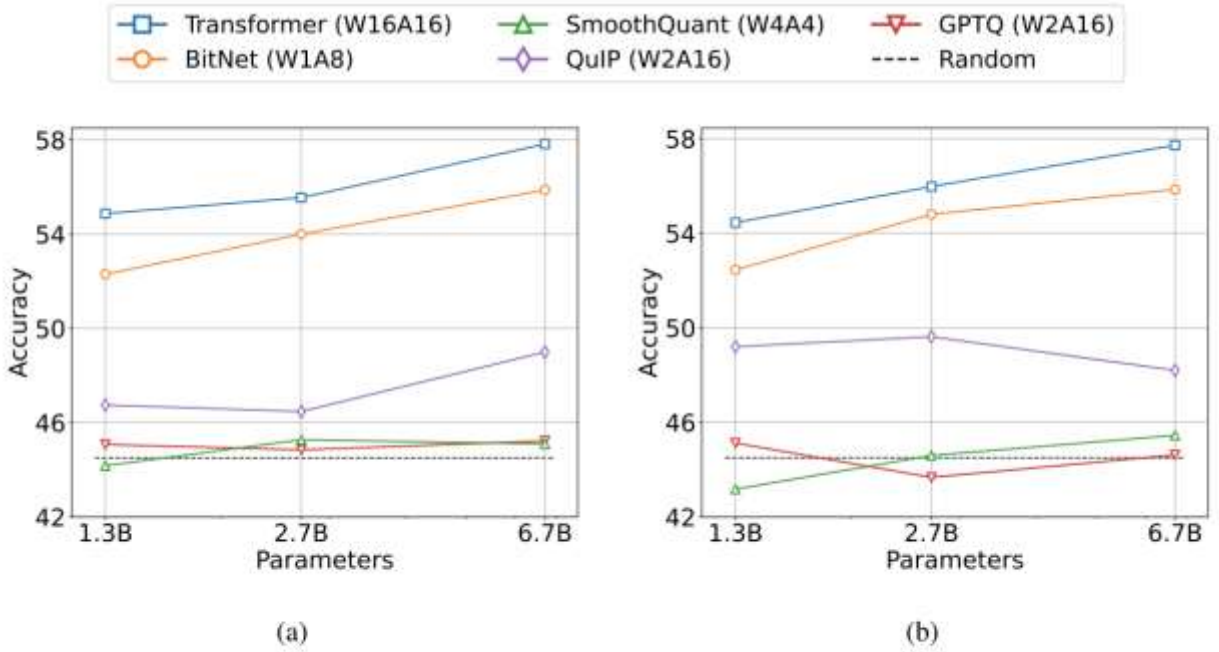
- BitNet achieves better performance than SoTA Post-Training Quant on the ultra-low bit models.

WBits	Methods	PTQ	PPL↓	WG↑	WGe↑	HS↑	SC↑	Avg↑
-	Random	✗	-	50.0	50.0	25.0	50.0	43.8
16	Transformer	✗	15.19	66.7	54.3	42.9	67.4	57.8
8	Absmax	✓	21.43	60.4	52.0	38.3	62.7	53.4
	SmoothQuant	✓	15.67	65.3	53.1	40.9	67.6	56.7
4	GPTQ	✓	16.05	57.2	51.2	39.9	63.4	52.9
	Absmax	✓	4.8e4	55.8	50.9	25.0	53.1	46.2
	SmoothQuant	✓	1.6e6	53.7	48.3	24.8	53.6	45.1
2	GPTQ	✓	1032	51.6	50.1	25.8	53.4	45.2
	QuIP	✓	70.43	56.1	51.2	30.3	58.4	49.0
1	Absmax	✓	3.5e23	49.8	50.0	24.8	53.6	44.6
	SmoothQuant	✓	3.3e21	50.5	49.5	24.6	53.1	44.4
1	**BitNet**	✗	17.07	66.3	51.4	38.9	66.9	55.9

Table 3: Zero-shot results for BitNet and the baselines (PTQ: Post-training quantization, WGe: Winogrande, WG: Winograd, SC: Storycloze, and HS: Hellaswag dataset).

07 BitNet vs SoTA PTQs

- BitNet achieves better performance than SoTA Post-Training Quant on the ultra-low bit models.



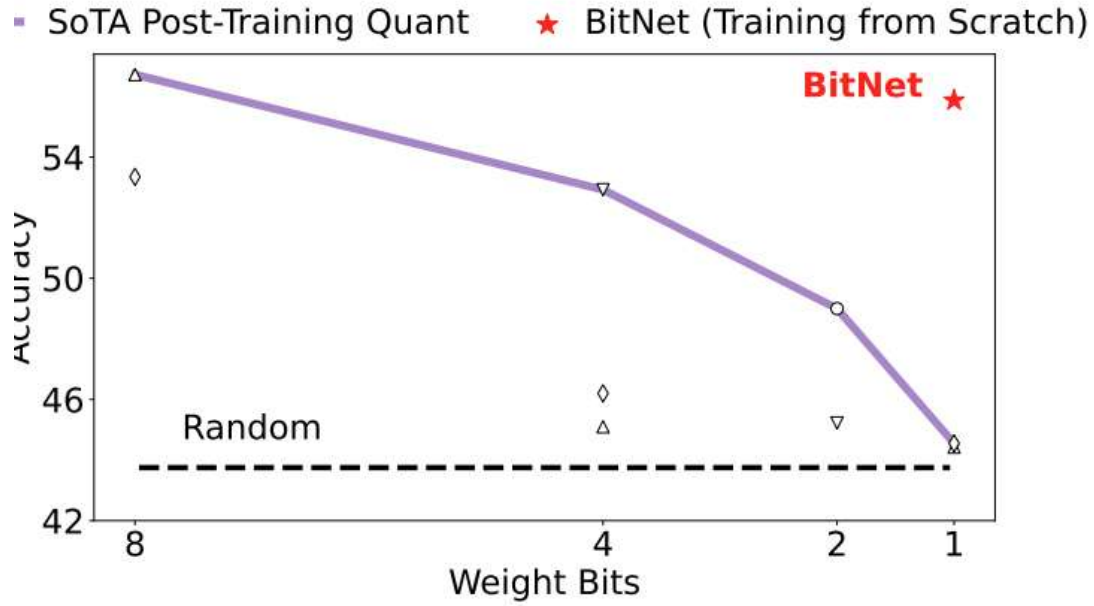


Figure 6: Zero-shot (Left) and few-shot (Right) results for BitNet and the post-training quantization baselines on downstream tasks.

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

Shuming Ma [*1], Hongyu Wang [*12], Lingxiao Ma [1], Lei Wang [1], Wenhui Wang [1],
Shaohan Huang [1], Li Dong [1], Ruiping Wang [2], Jilong Xue [1], Furu Wei [1]
[1] Microsoft Research, [2] University of Chinese Academy of Sciences

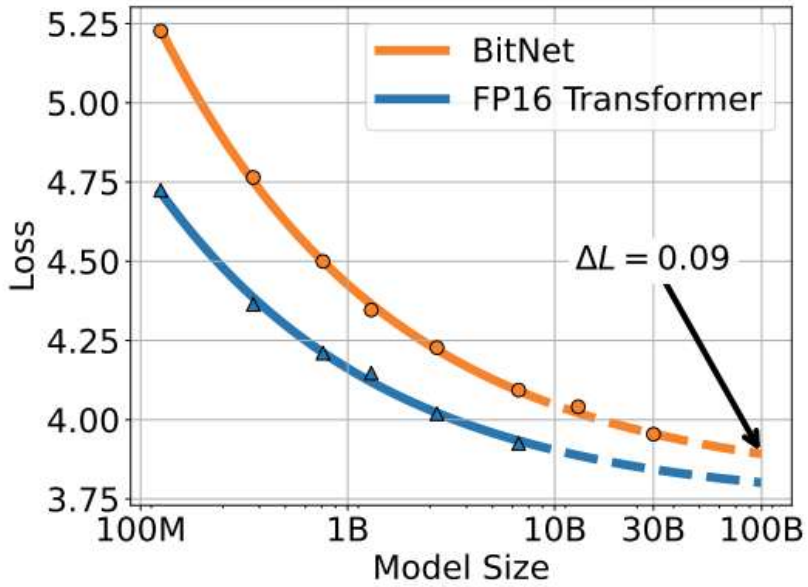
https://aka.ms/GeneralAI

Paper

BitNet b1.58

- Revisit the scaling law of BitNet b1
  - 😎 As the model size grows, the gap between 1-bit and FP16 LLM narrows
  - 😟 BitNet b1 requires a much large size (>100B) to match FP16 LLM

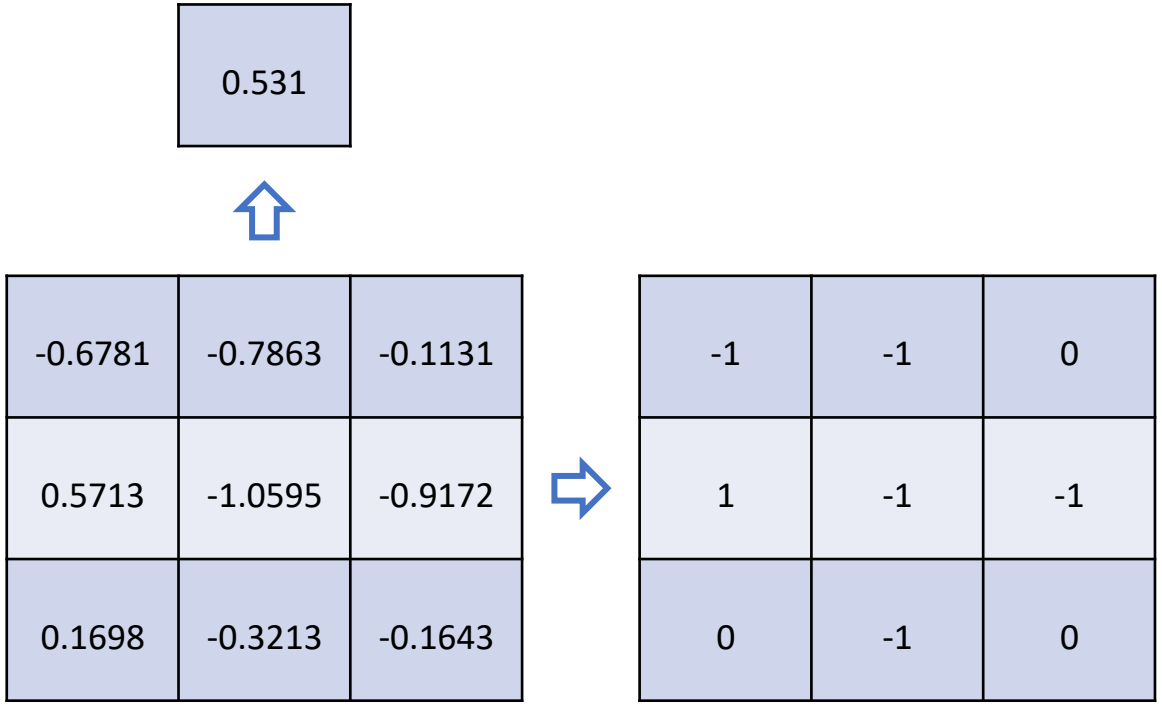
- Could BitNet match FP16 LLM under a smaller size?
  - Yes!
  - {-1, 1} -> {-1, 0, 1} significantly boosts the performance!



01 BitLinear

- Normalization: Sub-LN
- Weight quantization: 1.58-bit ($\log_2 3$)
  - Function: absmean

$$\widetilde{W} = \text{RoundClip}(\frac{W}{\gamma + \epsilon}, -1, 1), \qquad \gamma = \frac{1}{nm}\sum_{ij}|W_{ij}|.$$

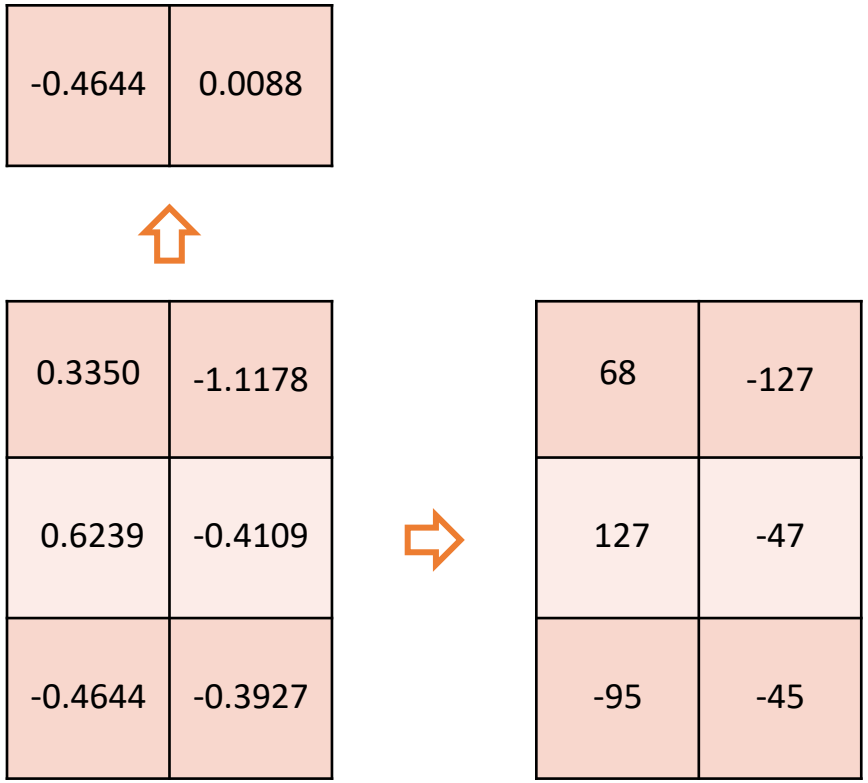


01 BitLinear

- Normalization: Sub-LN
- Weight quantization: 1.58-bit ($\log_2 3$)
  - Function: absmean

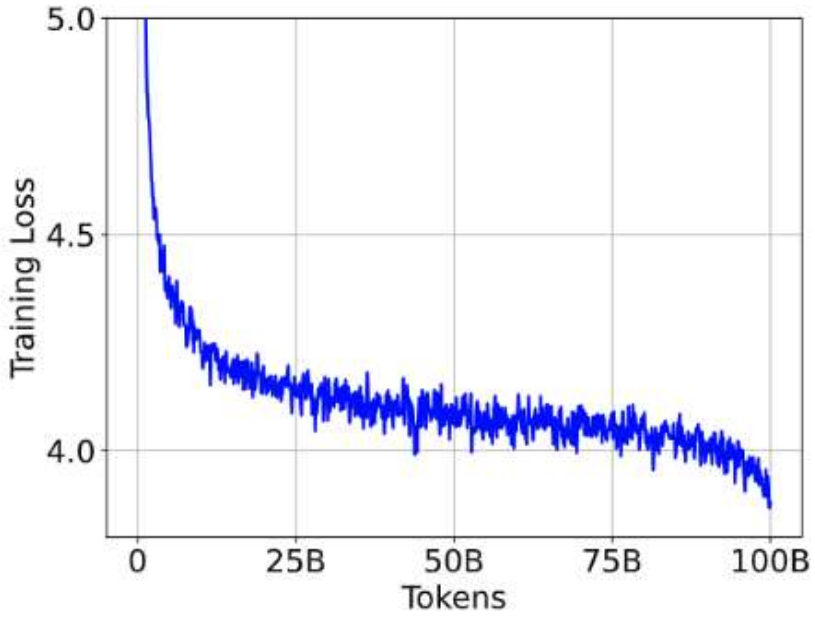
$$\widetilde{W} = \mathrm{RoundClip}(\frac{W}{\gamma+\epsilon}, -1, 1), \qquad \gamma = \frac{1}{nm}\sum_{ij}|W_{ij}|.$$

- Activation quantization: INT8
  - Function: symmetric absmax(per token)



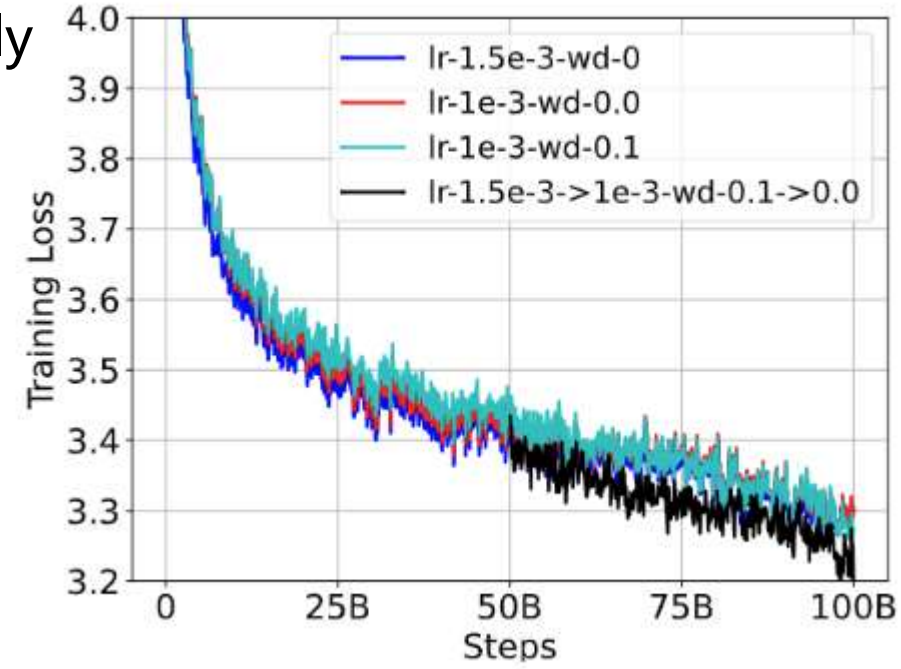
02 Training

- S-shape loss curve
  - The loss suddenly decreases at the end of training



02 Training

- S-shape loss curve
  - The loss suddenly decreases at the end of training
  - When wd=0, S-shape disappears; but it has slightly worse ppl than wd=0.1

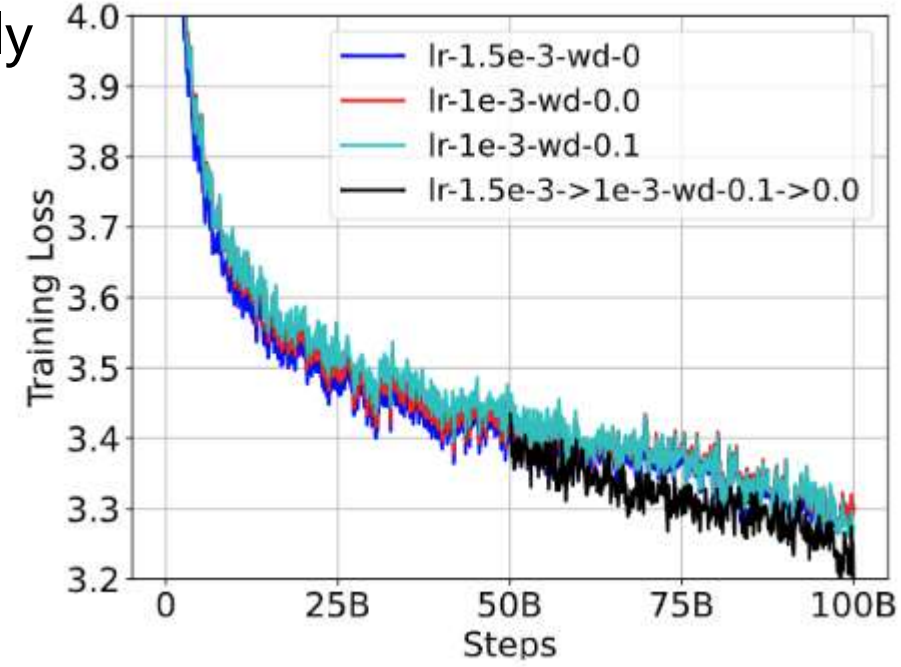


wd = weight decay

02 Training

- S-shape loss curve
  - The loss suddenly decreases at the end of training
  - When wd=0, S-shape disappears; but it has slightly worse ppl than wd=0.1

- Two-stage training
  - Stage-1: First half of training
    - High learning rate
    - Enabling weight decay
  - Stage-2: Second half of training
    - Lower learning rate
    - Disabling weight decay



wd = weight decay

02 Training

- **As the model size grows, the gap between 1.58-bit and FP16 narrows**

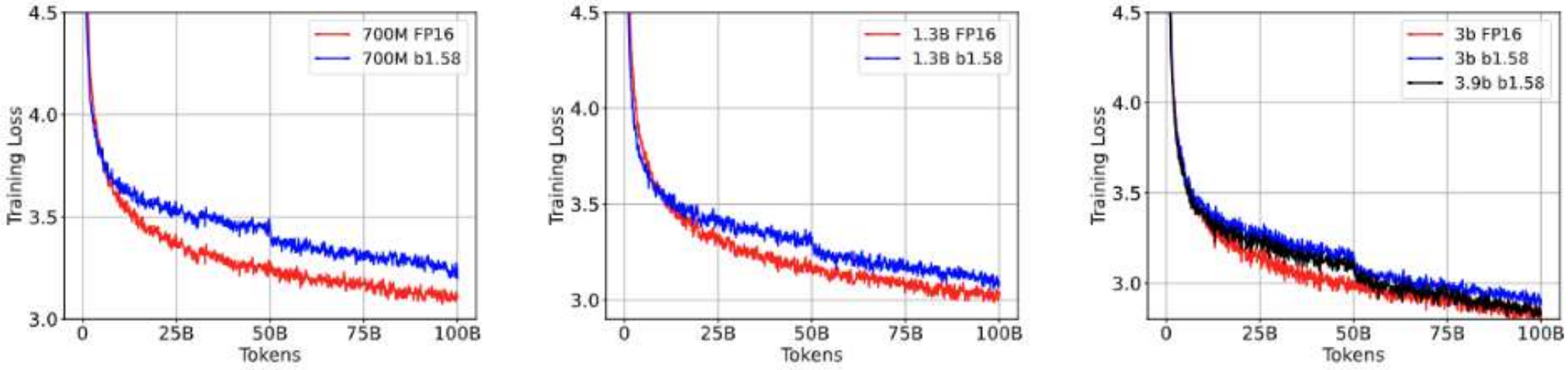


Figure 2: Training loss curves across different model sizes. The gap between full-precision models and BitNet b1.58 becomes narrower as the models scale.

03 Experiments

- **As the model size grows, the gap between 1.58-bit and FP16 narrows**
- **BitNet b1.58 matches FP16 LLM with 3B parameters and 100B tokens.**

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.

All models are trained with 100B tokens on Redpajama dataset.

03 Experiments

- **As the model size grows, the gap between 1.58-bit and FP16 narrows**
- **BitNet b1.58 matches FP16 LLM with 3B parameters and 100B tokens.**

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.

All models are trained with 100B tokens on Redpajama dataset.

04 Inference cost

- Implemented with an INT8 x INT2 kernel on 80G A100 cards.
- Compared with 70B LLaMA, BitNet b1.58 has:
  - 4x decoding Latency, only 14% memory consumption;

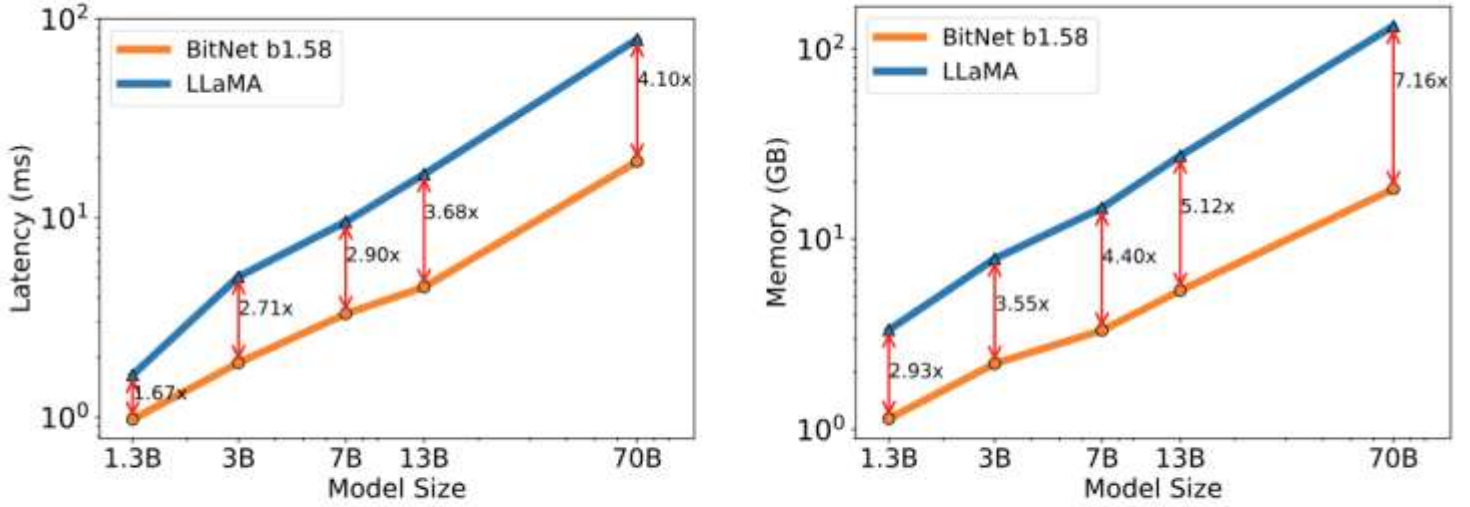


Figure 2: Decoding latency (Left) and memory consumption (Right) of BitNet b1.58 varying the model size.

https://github.com/microsoft/BitBLAS/tree/main/integration/BitNet

04 Inference cost

- Implemented with an INT8 x INT2 kernel on 80G A100 cards.
- Compared with 70B LLaMA, BitNet b1.58 has:
  - 4x decoding Latency, only 14% memory consumption;
  - 9x throughput, 11x maximum batch 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.

https://github.com/microsoft/BitBLAS/tree/main/integration/BitNet

04 Inference cost

- Implemented with an INT8 x INT2 kernel on 80G A100 cards.
- Compared with 70B LLaMA, BitNet b1.58 has:
  - 4x decoding Latency, only 14% memory consumption;
  - 9x throughput, 11x maximum batch size.

- New scaling law as for performance and inference cost:
  - 13B BitNet b1.58 is more efficient than 3B FP16 LLM;
  - 30B BitNet b1.58 is more efficient than 7B FP16 LLM;
  - **70B BitNet b1.58 is more efficient than 13B FP16 LLM.**

05 Scaling with more data

- 3B BitNet b1.58 with 2T tokens:

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.

The results of Stable-3B are from their reports.

BitNet b1.58 2B4T Technical Report

Shuming Ma*[1], Hongyu Wang*[12], Shanhan Huang[1], Xingxing Zhang[1]
Ying Hu[3], Ting Song[1], Yan Xia[1], Furu Wei[1]
[1] Microsoft Research, [2] University of Chinese Academy of Sciences,
[3] TsingHua University

https://aka.ms/GeneralAI

Paper

01 Scaling Native 1-bit LLM

- BitNet b1.58 2B – **the first native 1-bit LLM**
  - 1.58-bit weights & INT8 activations
  - Squared ReLU
  - Available at https://huggingface.co/microsoft/bitnet-b1.58-2B-4T
  - Inference framework (bitnet.cpp): https://github.com/microsoft/bitnet

- Pre-training
  - Data sources: DCLM, Fineweb-EDU, Synthetical math data, etc.
  - 4T tokens
  - **Two-stage lr & weight decay scheduling**

- SFT & DPO
  - Large learning rate
  - Training longer in SFT

02 Evaluation

- Comparable to Qwen2.5-1.5B on benchmarks
- Lower inference cost
  - Memory footprint: 0.4GB
  - CPU Latency*: 29ms
  - AOE energy: 0.028J

Benchmark (Metric)	LLaMA 3.2 1B	Gemma-3 1B	Qwen2.5 1.5B	SmolLM2 1.7B	MiniCPM 2B	BitNet b1.58 2B
Memory (Non-emb)	2GB	1.4GB	2.6GB	3.2GB	4.8GB	**0.4GB**
Latency (CPU; TPOT)	48ms	41ms	65ms	67ms	124ms	**29ms**
Energy (Estimated)	0.258J	0.186J	0.347J	0.425J	0.649J	**0.028J**
Training Tokens (Pre-training)	9T (pruning & distillation)	2T (distillation)	18T	11T	1.1T	4T
ARC-Challange (0-shot; Acc,norm)	37.80	38.40	46.67	43.52	44.80	**49.91**
ARC-Easy (0-shot; Acc,norm)	63.17	63.13	**76.01**	62.92	72.14	74.79
OpenbookQA (0-shot; Acc,norm)	34.80	38.80	40.80	**46.00**	40.20	41.60
BoolQ (0-shot; Acc)	64.65	74.22	78.04	75.78	**80.67**	80.18
HellaSwag (0-shot; Acc,norm)	60.80	57.69	68.28	**71.71**	70.81	68.44
PIQA (0-shot; Acc,norm)	74.21	71.93	76.12	76.12	76.66	**77.09**
WinoGrande (0-shot; Acc)	59.51	58.48	62.83	68.98	61.80	**71.90**
CommonsenseQA (10-shot; Acc)	58.48	42.10	**76.41**	63.55	71.74	71.58
TruthfulQA (10-shot; MC2)	43.80	38.66	**46.67**	39.90	41.41	45.31
TriviaQA (5-shot; EM)	37.60	23.49	38.37	**45.97**	34.13	33.57
MMLU (5-shot; Acc)	45.58	39.91	**60.25**	49.24	51.82	53.17
HumanEval+ (0-shot; Pass@1)	31.10	37.20	**50.60**	28.00	43.90	38.40
GSM8K (4-shot; EM)	38.21	31.16	56.79	45.11	4.40	**58.38**
MATH-500 (0-shot; EM)	23.00	42.00	**53.00**	17.60	14.80	43.40
IFEval (0-shot; Instruct-Strict)	62.71	**66.67**	50.12	57.91	36.81	53.48
MT-bench (0-shot; Average)	5.43	6.40	6.12	5.50	**6.57**	5.85
Average	44.90	43.74	**55.23**	48.70	42.05	54.19

Table 1: Comparison of BitNet b1.58 2B4T with leading open-weight full-precision LLMs of similar size (1B-2B parameters) on efficiency metrics and performance across a wide range of benchmarks. All models compared are instruction-tuned versions.

*For each model, we generated 128 tokens and report the average latency per token on a Surface Laptop Studio 2 powered by a 13th Gen Intel Core i7-13800H processor

02 Evaluation

- Comparable to Qwen2.5-1.5B on benchmarks
- Lower inference cost
  - Memory footprint: 0.4GB
  - CPU Latency*: 29ms
  - AOE energy: 0.028J
- Outperforms Qwen2.5-1.5B (INT4)
  - PTQ leads to significant loss on Math task

Benchmark (Metric)	Qwen2.5 1.5B-bf16	Qwen2.5 1.5B-GPTQ-int4	Qwen2.5 1.5B-AWQ-int4	BitNet b1.58 2B
Memory (Non-emb)	2.6GB	0.7GB	0.7GB	0.4GB
Activation	bf16	bf16	bf16	int8
MMLU (5-shot; Acc)	**60.25**	58.06	57.43	53.17
GSM8K (4-shot; EM)	56.79	50.57	50.64	**58.38**
IFEval (0-shot; Instruct-Strict)	50.12	47.84	45.44	**53.48**
Average	**55.72**	52.15	51.17	55.01

Table 2: Comparison of BitNet b1.58 (2B) against Qwen2.5 1.5B in its original bf16 precision and after INT4 post-training quantization (GPTQ and AWQ). All models shown are based on instruction-tuned checkpoints.

02 Evaluation

- Comparable to Qwen2.5-1.5B on benchmarks
- Lower inference cost
  - Memory footprint: 0.4GB
  - CPU Latency*: 29ms
  - AOE energy: 0.028J
- Outperforms Qwen2.5-1.5B (INT4)
  - PTQ leads to significant loss on Math task
- Pre-training is essential for 1-bit models

Benchmark (Metric)	Bonsai 0.5B	OLMo-Bitnet 1B	Falcon3-1.58bit 7B	Llama3-8B-1.58 8B	BitNet b1.58 2B
Native 1-bit	✓	✓	✗	✗	✓
ARC-Challange (0-shot; Acc,norm)	33.19	26.54	37.80	43.69	**49.91**
ARC-Easy (0-shot; Acc,norm)	58.25	25.38	65.03	70.71	**74.79**
OpenbookQA (0-shot; Acc,norm)	33.60	28.20	38.20	37.20	**41.60**
BoolQ (0-shot; Acc)	58.44	52.48	72.14	68.38	**80.18**
HellaSwag (0-shot; Acc,norm)	48.01	25.88	59.46	**68.56**	68.44
PIQA (0-shot; Acc,norm)	70.02	50.49	72.36	75.30	**77.09**
WinoGrande (0-shot; Acc)	54.46	51.54	60.14	60.93	**71.90**
CommonsenseQA (10-shot; Acc)	18.43	19.49	67.08	28.50	**71.58**
TruthfulQA (10-shot; MC2)	40.65	**49.05**	43.29	39.13	45.31
TriviaQA (5-shot; EM)	10.84	0.00	0.00	19.82	**33.57**
MMLU (5-shot; Acc)	25.74	25.47	42.79	35.04	**53.17**
Average	41.06	32.22	50.76	49.75	**60.68**

Table 3: Performance comparison of BitNet b1.58 2B4T against other open-weight 1-bit models. This includes natively trained 1-bit models (Bonsai-0.5B, OLMo-Bitnet-1B) and larger models post-training quantized to 1.58-bit (Falcon3-1.58bit-7B, Llama3-8B-1.58).

BitNet v2: Native 4-bit Activations with Hadamard Transformation for 1-bit LLMs

Hongyu Wang*[12], Shuming Ma *[1], Furu Wei [1]

[1] Microsoft Research, [2] University of Chinese Academy of Sciences

https://aka.ms/GeneralAI

Paper

01 Paradigm shift in 1-bit LLMs

- **Memory-bound -> Compute-bound**

- Activation Sparsification
  - **Settings: single/small-batch**
  - Expert-level (MoE), FFN-level (PowerInfer), Linear-level (Q-Sparse)

- Activation Quantization
  - **Settings: large-batch**
  - 4-bit activation

Exponential scaling law with regards to activation sparsity

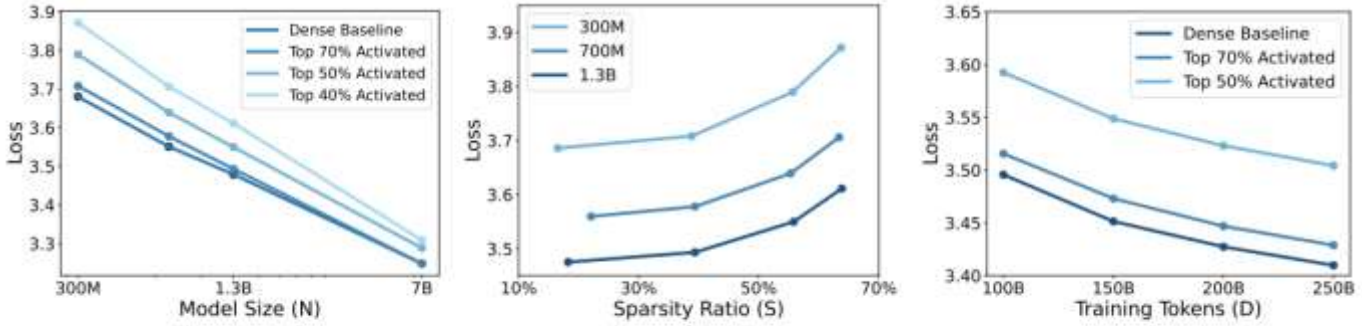
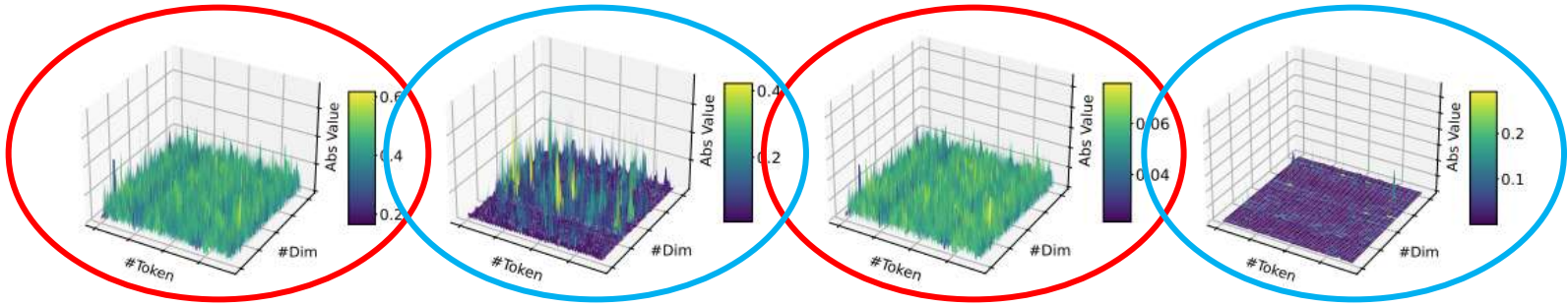


Figure 4: The scaling curves of the sparsely-activated models regrading to the model size given a fixed sparsity ratio S (Left), and regrading to the sparsity ratio given a fixed model size N (Right).

02 Challenge

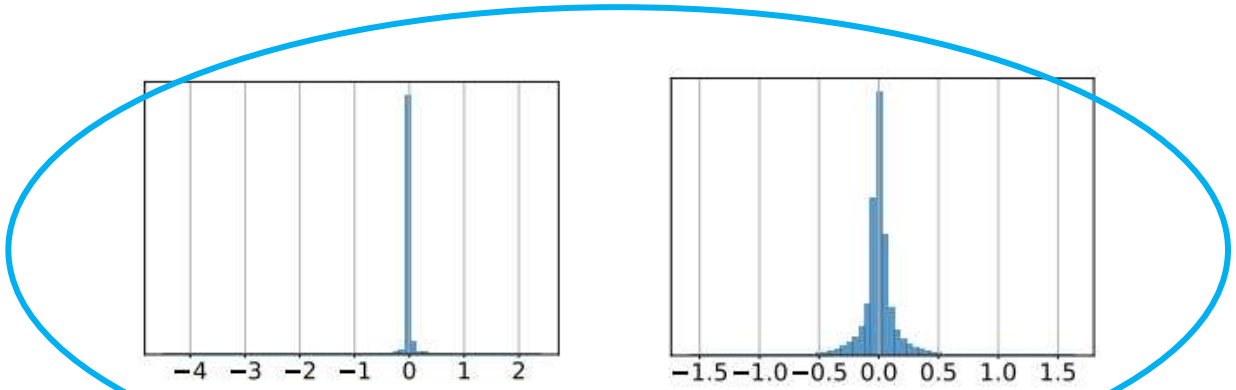
- Distinct distributions in LLMs
- Outliers emerges
  - as the model size grows
  - as the training progresses
- Challenges posed by outliers
  - Down proj is very sensitive to INT4 Act
  - **Gradient approximation**

Gaussian-like shape, suitable for quantization!



(a) $\mathbf{W}_{qkv}$ of BitNet b1.58 (b) $\mathbf{W}_{o}$ of BitNet b1.58 (c) $\mathbf{W}_{up,gate}$ of BitNet b1.58 (d) $\mathbf{W}_{down}$ of BitNet b1.58

Outliers, suitable for sparsification



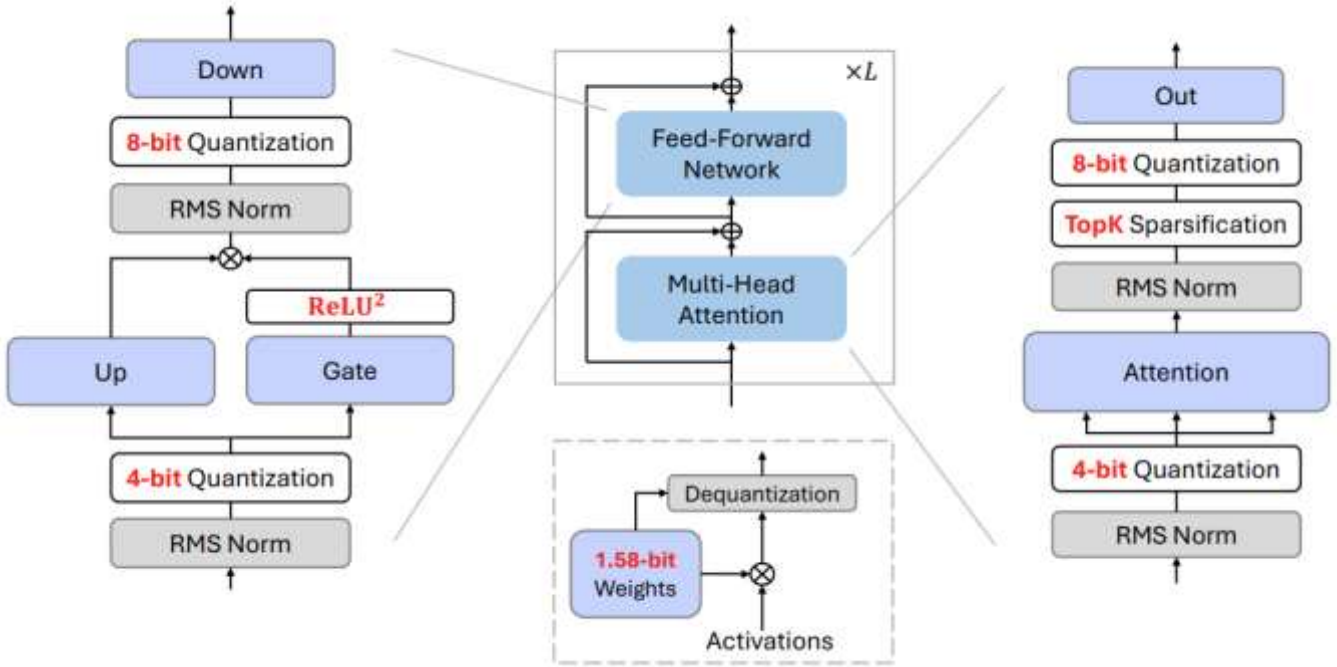
(a) $\mathbf{W}_{down}$ of BitNet b1.58 (c) $\mathbf{W}_{o}$ of BitNet b1.58

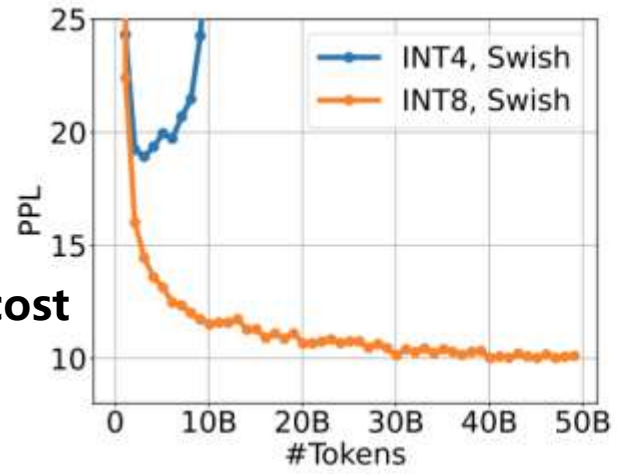
Paper

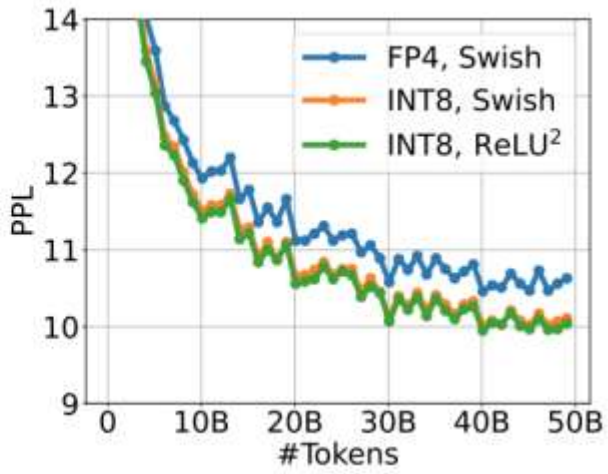
02 Challenge

- Distinct distributions in LLMs
- Outliers emerges
  - as the model size grows
  - as the training progresses
- Challenges posed by outliers
  - Down proj is very sensitive to INT4 Act
  - **Gradient approximation**
- **A direct solution: BitNet a4.8**
  - **Hybrid quantization and sparsification**
  - **Cont-train from BitNet b1.58 with small cost**
  - **Act. quantizer: absmean**

Paper







02 Challenge

- Distinct distributions in LLMs
- Outliers emerges
  - as the model size grows
  - as the training progresses
- Challenges posed by outliers
  - Down proj is very sensitive to INT4 Act
  - **Gradient approximation**
- **A direct solution: BitNet a4.8**
  - **Hybrid quantization and sparsification**
  - **Cont-train from BitNet b1.58 with small cost**
  - **Act. quantizer: absmean**

Paper

Models	Size	PPL↓	ARCc↑	ARCe↑	HS↑	PQ↑	WGe↑	Avg↑
LLaMA LLM	700M	11.44	27.13	43.27	44.70	68.12	53.99	47.44
BitNet b1.58		12.32	25.00	42.68	42.08	66.97	54.14	46.17
BitNet a4.8 (FP4)		12.40	25.17	42.68	42.36	66.27	52.96	45.89
BitNet a4.8		12.40	25.17	41.58	42.44	66.38	53.04	45.72
LLaMA LLM	1.3B	10.82	27.90	45.16	47.65	69.91	53.35	48.79
BitNet b1.58		11.27	27.65	45.33	46.86	68.39	54.06	48.46
BitNet a4.8 (FP4)		11.38	28.50	44.36	47.03	68.61	54.06	48.51
BitNet a4.8		11.35	28.50	44.15	46.98	68.34	54.14	48.42
LLaMA LLM	3B	9.61	29.95	48.11	55.25	71.76	57.46	52.51
BitNet b1.58		9.97	29.27	49.41	54.42	70.89	57.54	52.30
BitNet a4.8 (FP4)		9.99	29.10	49.24	54.60	71.38	56.12	52.08
BitNet a4.8		9.97	28.33	49.58	54.62	71.16	54.38	51.61
LLaMA LLM	7B	9.20	33.36	51.22	58.33	73.34	58.41	54.93
BitNet b1.58		9.24	32.00	50.88	59.79	72.96	59.83	55.09
BitNet a4.8 (FP4)		9.42	31.57	51.22	58.20	72.47	59.59	54.61
BitNet a4.8		9.37	31.66	50.88	58.78	73.01	59.35	54.74

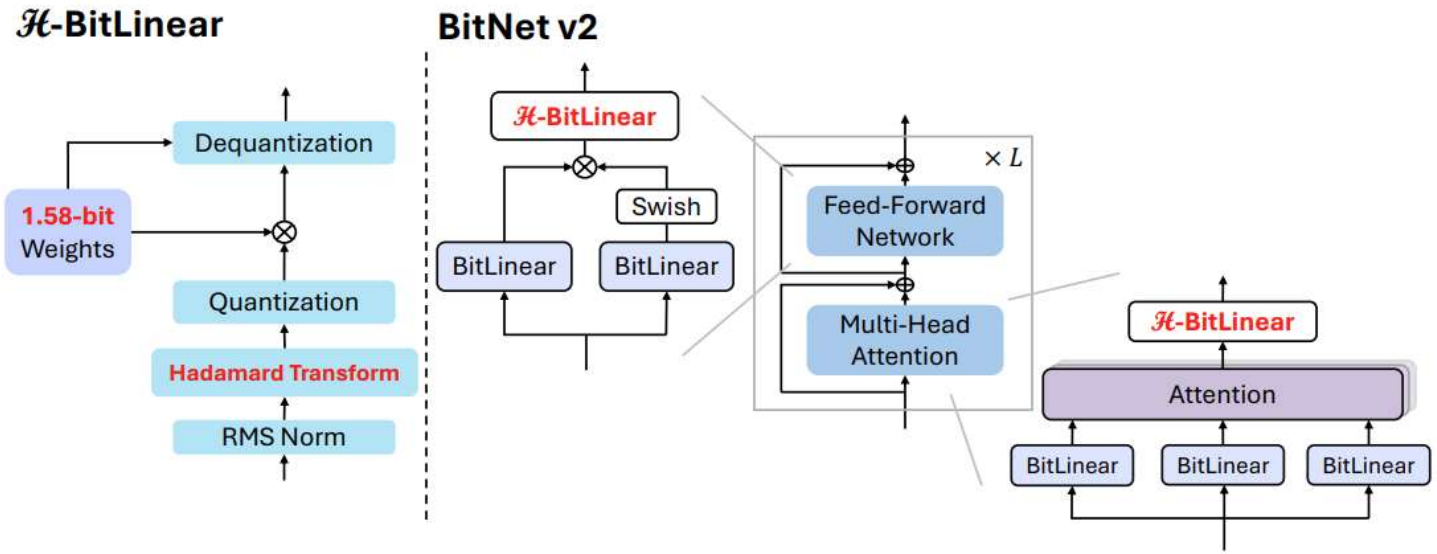
Table 1: Perplexity and results of BitNet a4.8, BitNet b1.58 and LLaMA LLM on the end tasks. The standard variance of error for average scores is 1.06%.

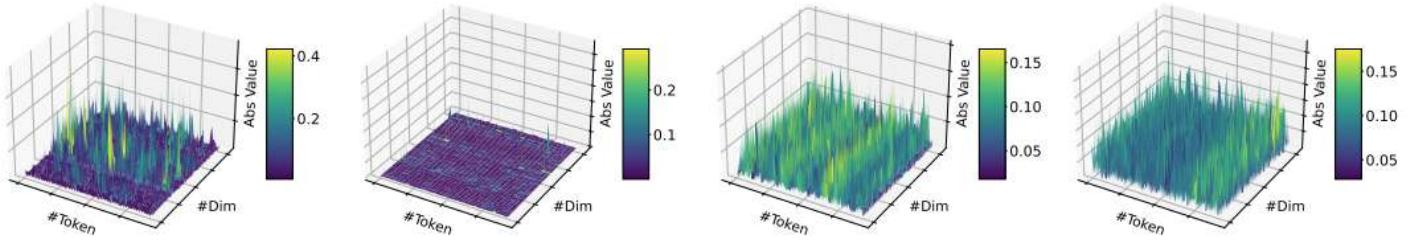
Models	Activated	QKV	Out	Up	Gate	Down	Overall
LLaMA LLM	679M	0.0	0.0	0.0	0.0	0.0	0.0
BitNet b1.58	638M	1.2	5.9	1.2	1.2	21.8	6.2
BitNet a4.8	390M	12.1	50.0	66.2	12.1	80.9	42.5
LLaMA LLM	1.2B	0.0	0.0	0.0	0.0	0.0	0.0
BitNet b1.58	1.1B	1.3	5.8	1.2	1.2	22.8	6.4
BitNet a4.8	0.7B	12.0	50.0	65.9	12.1	81.8	42.7
LLaMA LLM	3.2B	0.0	0.0	0.0	0.0	0.0	0.0
BitNet b1.58	3.0B	1.4	7.1	1.3	1.3	30.0	8.2
BitNet a4.8	1.8B	12.1	50.0	70.7	12.1	85.6	44.7
LLaMA LLM	6.5B	0.0	0.0	0.0	0.0	0.0	0.0
BitNet b1.58	6.0B	1.7	11.2	1.4	1.4	24.2	7.3
BitNet a4.8	3.4B	12.1	50.0	71.4	12.0	84.2	44.5

Table 2: Detailed sparsity of BitNet a4.8, BitNet b1.58 and LLaMA LLM on the valid set of C4.

03 BitNet v2: Native 4-bit Act. for 1-bit LLM

- H-BitLinear
  - Add **Hadamard Transform** before activation quantization
- BitNet v2
  - Down proj, o_proj use H-BitLinear
  - Other projs use BitLinear
- Training
  - INT8 training, absmax quantizer (~95%)
  - INT4 training, absmean quantizer (~5%)





(a) $\mathbf{W}_o$ of BitNet b1.58 (b) $\mathbf{W}_{down}$ of BitNet b1.58 (c) $\mathbf{W}_o$ of BitNet v2 (d) $\mathbf{W}_{down}$ of BitNet v2

03 Experiments

- BitNet v2 (a8) matches BitNet b1.58,

BitNet v2 (a4) matches BitNet a4.8

Models	Size	PPL↓	ARCc↑	ARCe↑	HS↑	PQ↑	WGe↑	LBA↑	Avg↑
BitNet b1.58	400M	13.37	24.32	43.01	39.51	64.91	51.93	45.51	44.87
BitNet a4.8		13.61	24.15	41.75	39.48	65.18	53.59	44.34	44.75
BitNet v2 (a8)		13.50	23.29	43.06	39.06	64.74	50.59	45.26	44.33
BitNet v2 (a4)		13.78	23.29	41.46	38.33	65.45	50.59	44.56	43.95
BitNet b1.58	1.3B	11.02	27.90	49.58	48.85	69.80	55.80	54.12	51.01
BitNet a4.8		11.15	27.47	49.20	48.72	69.64	56.51	53.85	50.90
BitNet v2 (a8)		11.14	27.90	49.96	48.37	69.42	57.22	54.14	51.17
BitNet v2 (a4)		11.33	27.56	49.58	48.00	68.23	55.49	53.58	50.41
BitNet b1.58	3B	9.71	28.84	54.80	56.39	71.44	59.35	60.47	55.22
BitNet a4.8		9.80	29.01	55.01	55.92	71.76	59.59	59.85	55.19
BitNet v2 (a8)		9.72	30.55	55.56	57.19	71.33	58.72	60.90	55.71
BitNet v2 (a4)		9.85	28.92	55.01	56.59	71.65	59.67	60.74	55.43
BitNet b1.58	7B	9.09	31.74	59.51	61.49	74.37	59.98	61.63	58.12
BitNet a4.8		9.16	31.91	59.09	61.06	74.16	59.67	61.54	57.91
BitNet v2 (a8)		9.14	32.94	58.54	61.08	74.10	61.48	64.22	58.73
BitNet v2 (a4)		9.24	32.42	58.00	60.71	74.27	60.85	63.52	58.30

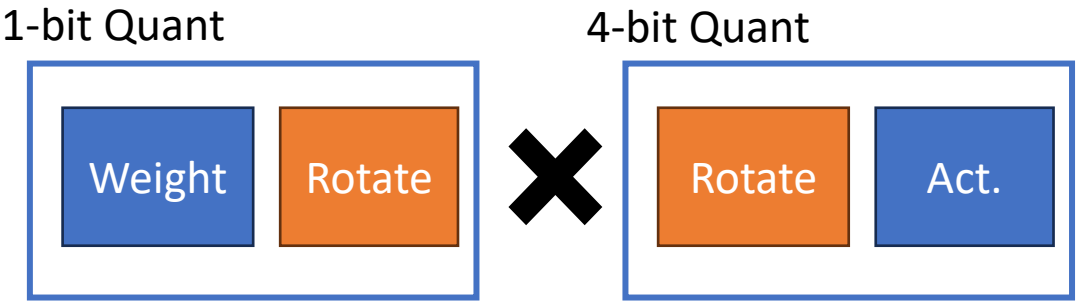
Table 1: Perplexity and results of BitNet v2, BitNet a4.8 and BitNet b1.58 on the end tasks.

03 Experiments

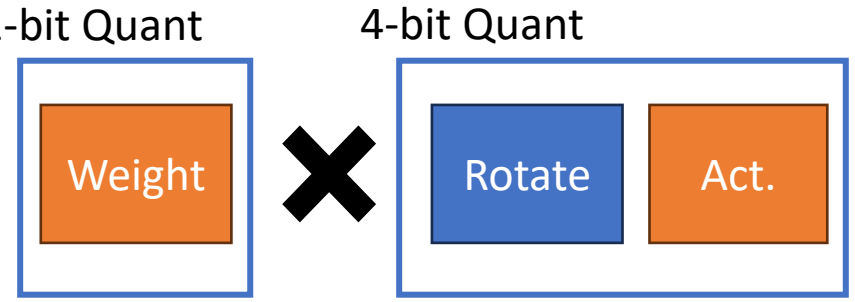
- BitNet v2 (a8) matches BitNet b1.58, BitNet v2 (a4) matches BitNet a4.8

- Comparison with PTQ
  - 1.58-bit weights are sensitive to PTQ

PTQ:

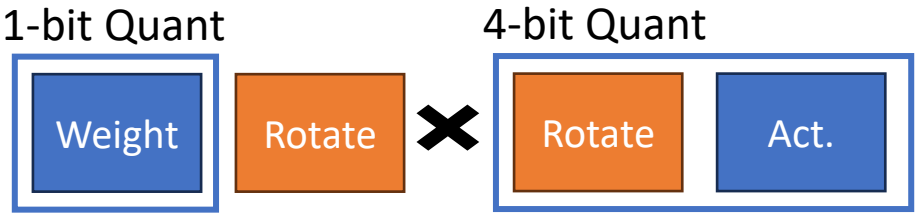


BitNet v2:



Models	PPL↓	ARCc↑	ARCe↑	HS↑	PQ↑	WGe↑	LBA↑	Avg↑
w/o fusing rotary matrix to $W_{qkv,up,gate}$								
QuaRot *W16-A4*	13.52	26.28	47.43	45.92	65.89	51.46	42.34	46.55
SpinQuant	13.52	25.60	47.35	45.52	67.25	52.49	42.52	46.79
QuaRot *W1.58-A4*	20.83	24.74	40.78	40.54	62.89	49.33	36.89	42.53
SpinQuant	19.80	24.74	40.19	40.77	62.73	52.09	39.24	43.29
BitNet v2 (a4)	**11.33**	**27.56**	**49.58**	**48.00**	**68.23**	**55.49**	**53.58**	**50.41**

Table 4: Perplexity and zero-shot accuracy of BitNet v2, QuaRot and SpinQuant on the end tasks.



03 Experiments

- BitNet v2 (a8) matches BitNet b1.58,

BitNet v2 (a4) matches BitNet a4.8

- Comparison with PTQ
  - 1.58-bit weights are sensitive to PTQ
- Ablations
  - Weight rotation is unnecessary

Methods	#Bits	1.3B		3B	
		Acc.↑	PPL↓	Acc.↑	PPL↓
No rotation	W1.58A8	diverged		diverged	
Weight & activation rotation		50.47	11.14	55.55	9.69
Activation rotation		51.16	11.14	55.71	9.72
No rotation	W1.58A4	diverged		diverged	
Weight & activation rotation		50.09	11.33	54.98	9.81
Activation rotation		50.41	11.33	55.43	9.85

Table 5: Ablations on the Hadamard transformation of $\mathcal{H}$-BitLinear across various sizes.

1-bit LLM Family

- Models
  - BitNet b1.58 2B4T (by Microsoft)
  - Falcon-3 and Falcon-E (by TII)
  - Llama3-8B-1.58-100B-tokens (by HuggingFace)
  - ...
- Inference framework
  - BitNet.cpp (GPU/CPU)
  - BitBLAS (GPU)
  - T-MAC (CPU)
- Hardware
  - Slim-Llama

BitNet b1 paper

BitNet b1.58 paper

BitNet b1.58: Tips, Code and FAQ

BitNet a4.8 paper

BitNet v2 paper

Q-Sparse paper

Takeaway

① **As the model size grows, the gap between 1-bit and FP16 LLM narrows.**

② **BitNet matches FP16 LLM staring from 3B parameters and 100B tokens.**

③ **New scaling law as for performance and inference cost.**
- **70B BitNet b1.58 is more efficient than 13B FP16 LLaMA (latency, memory, energy)**

④ **1-bit models open the door to Next-Gen hardware for LLM.**

⑤ **Scaling laws closely tied to specific architecture and its training recipe. Optimization is critical for low-bit models!**