

arXiv:2411.12364v2 [cs.LG] 6 Feb 2025

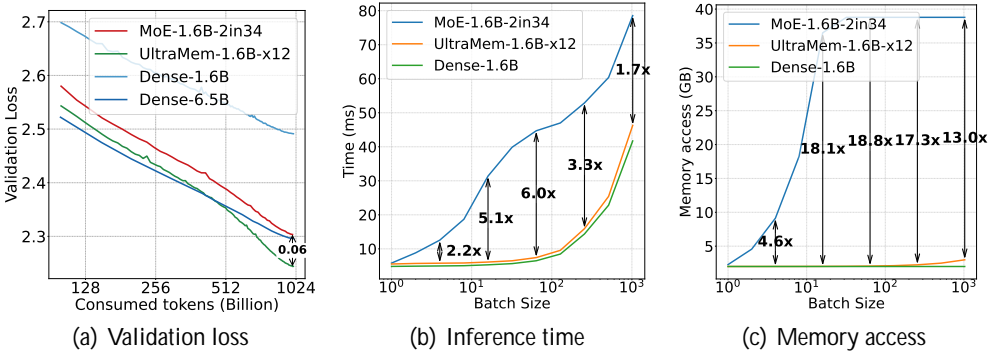
Ultra-Sparse

Zihao Huang, Qiyang Min, Hongzhi Huang, Defa Zhu, Yutao Zeng, Ran Guo, Xun Zhou
Seed-Foundation-Model Team, ByteDance
fhuangzi hao. notabot, mi nqi yang, huanghongzhi . 51, zhudefa, yutao. zeng, guoran. 94, zhouxung@bytedance. com

ABSTRACT

computation MOE MOE
UltraMem
the tharem thar ultramem 2000 MOE

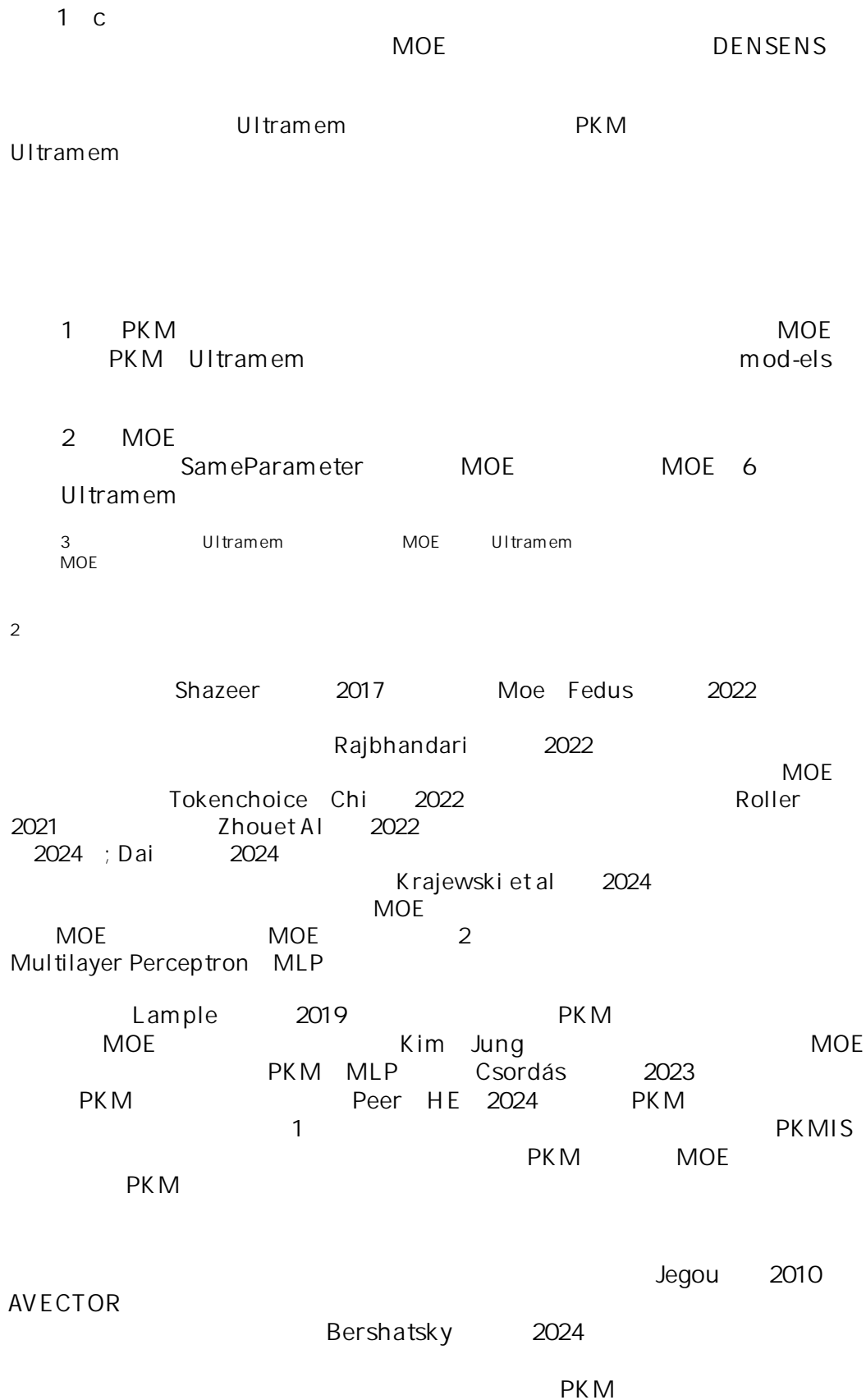
1

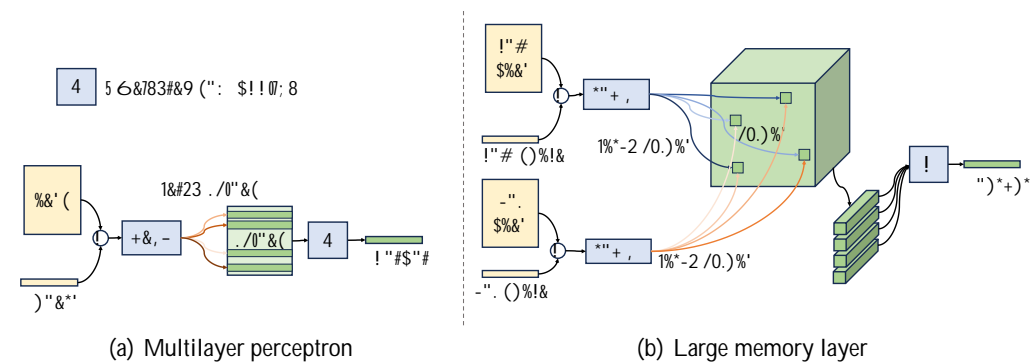


1 X b c MOE Ultramem 1
/ Cachelengths 2048 A -SXM- GB

LLMS NLP Radford
2019; Brown 2020
Fedus 2022; Jiang 2024 PKM MOE
2019 MOE Lample
PKM MOE

1 b 2 6 2 6





2 MLP
M MLP
2-Doogle MLP 1-D

3 ULTRAMEM

3.1 PRELIMINARY

2019

An external

LML

PKM

Lample

2 b

$\mathbf{q} \in \mathbb{R}^d$

$\mathbf{K} \in \mathbb{R}^{n \times d}$

$\mathbf{v} \in \mathbb{R}^{n \times d}$

$\mathbf{s} = (\mathbf{K}\mathbf{q})$

$\mathbf{o} = \mathbf{V}^T \mathbf{s}$

(1)

s

o

MLP

gelu

Geva

2020

MLP

2 a

$n > 10^6$

M

1

TOP-M

$n \times n$

$n =$

2-D

2D

address

i, j

2 b

$n \times i + j.j.j.j.$

$$\mathbf{s}_{row} = \text{TopM}(\mathbf{K}_{row}q_{row}(\mathbf{x})); \quad \mathbf{s}_{col} = \text{TopM}(\mathbf{K}_{col}q_{col}(\mathbf{x})); \quad (2)$$

$$\text{s_grid} = \text{topm} \cdot \text{s_row} + \text{s_col} \quad \text{o} = \text{v} \times \text{softmax}(\text{vec}(\text{s_grid})) \quad 3 \quad \text{oftMax}(\text{vec}(\mathbf{S}_{grid})); \quad (3)$$

$$k_{topm} = \frac{Rn \times d \times k}{2} \times \frac{q \times col}{r \times di} \times \frac{r \times d \times k}{x} \times Rdi \text{ to row and column}$$

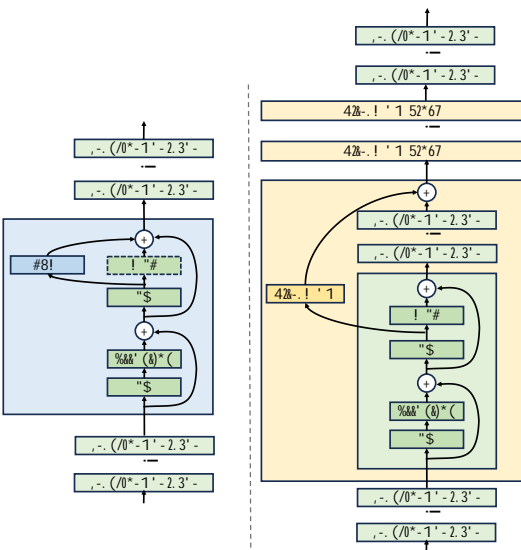
TOP-M S S col M S The last Top-M
n TOP-M O n log M M 2 n + m 2 n log M
S PKM
h
pkmheads

3.2

PKM PKM Minor
1 3 SoftMax Shen
2023; Csordás 2023 2 LN BA
4 PKM
Allealayer Howard 2017 5 5 5
Ainslie 2023 deformanceImpact 6 D V
HidDendimension

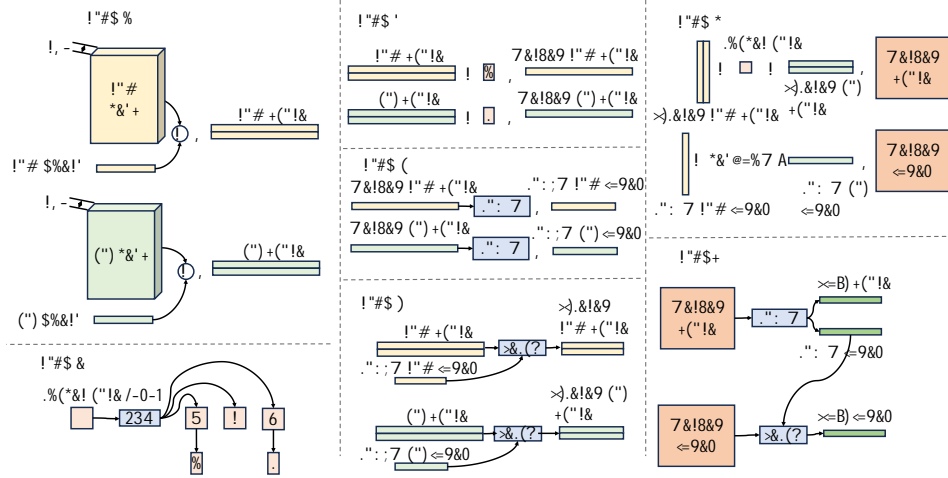
Ultramem 3

Apre-Layernorm MLP PKM Kim Jung 2020 PKM PKM



1 n
2 i j TOP-
J J Top- Top-M I
3 GPU
GPU
1 3
ThetransFormer ory Mem -

3 PKM Ultramem



4

TDQKR

$r = 2$ " "

TDQKR

1 2 TDQKR Malik Becker 2018 TDQKR
4 Tucker rank-R

$$S_{row} = K_{row} q_{row}(x); \quad S_{col} = K_{col} q_{col}(x); \quad (4)$$

$$S_{grid} = \text{TopM}(S_{row}^> \quad C \quad S_{col}); \quad (5)$$

s scol $R \times n$ c $R \times r$ Tucker Core
N \times R Thequery Key K
Kcol $R \times N \times D$ K /R Q Qcol $R \times D$ k /r 4 1

5

Top-M

Top-M

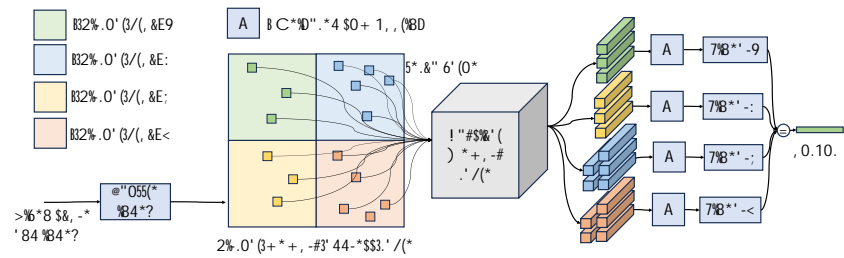
dorank-

$$C \quad ut^>; \quad \text{TopM}(S_{row}^> \quad C \quad S_{col}) \quad \text{TopM}((u^> S_{row})^> \quad (t^> S_{col})) \quad (6)$$

u t $R \times M$ Objective topm T scol $R \times n$
S Grid M TOP-M Outnon-Top SteelIndex 4 3

$$c \quad ut \quad 7 \quad srow = i \quad (7) \quad S_{col} = |_{\text{TopM}}(t^> S_{col}) \quad S_{col} \quad (9)$$

s grid = topm u srow s row \times c x scol 10 i topm .
top-m 7 10 1 4 5 8 9 4 Step
2007 u tucker 4 2 theleading



5 IVE e = 4 m = 16 ive

C = U T SVD 11 (11)

$$L_{aux} = \frac{1}{r} \sum_{i=2}^r (\max(0, \lambda_i)) ^2 ; \tag{12}$$

C C C Rank-

IVE

Avirtual E > 1 E

reparamemeterization {w p |p [e] wp Rdv× d'v} pth

Virtual Memory Block P

$$\mathbf{V}_p = \mathbf{V}\mathbf{W}_p; \tag{13}$$

e] d'v

$$= \begin{bmatrix} 0 & 1 \\ & d\ v \end{bmatrix}$$

V Rowand

nsion of physical

E · n · D V · D 'V E Times GPU

i j p i j address p

sum pliming pliming 3

$$\hat{s} = \text{Shuffle}(\text{vec}(\mathbf{S}_{grid})); \tag{14}$$

$$\mathbf{o} = \mathbf{V}^> \quad \hat{s} = \sum_p \mathbf{V}_p^> \quad \hat{s}_p = \sum_p \mathbf{W}_p^> \mathbf{V}^> \hat{s}_p \tag{15}$$

P p-th 15

from e-n-d v-d'v e-b-d v-d'v b
noextra gpu 5 IVE

MCS PKM D V
Tucker Core C C =

c i s {c i } hi = 1

aps $S^{(i)}_{tucker = s_{row} c_i s_{col}}$ Obviously,

$$s_{tucker = s_{row}} \sum_i C^{(i)} S_{col} = \sum_i S_{row}^{>} C^{(i)} S_{col} = \sum_i S^{(i)} \quad (16)$$

Tucker TOP-M S Tucker $V = [V^{(1)} \dots V^{(h)}]$ $R \times d \times h$ (i) on

$$o = [s^{(1)} > V^{(1)}; \dots; s^{(h)} > V^{(h)}] > \quad (17)$$

IVE 15

PKM n dv
1/D V LMLSH PKM SoftMax
MLP MLP
n L Brown 2020 L
e MHI m n h e expexpansion navith n
Ultramem n L Top-M 1
A

4 MOE
" "

2inn MoE d MLP 4D B MOE
2D 2 MOE / Anexample min b n
moe x d 2 Ultramem D/ TOP-M
BM N x D/ MOE
Ultramem
MOE

1 2in MOE x 12 1 Ultramem 16
64 7 tomoe Ultramem

5
Ultramem Ultramem Moe
TOP-M

1 Ultramem 12 2in MOE this

5.1

Llama

Touvron

2020

Redpajama

2023

CommonCrawl Web C

1

Redpajama

C

Redpajama

Ra el

Tokenizer

GPT-Neox

Black

2022

50,432

BPE

Sennrich

2015

MMLU

Trivia-QA

GPOA

ARC

HELLASWAG

WINOGRANDE

Agieval

seeapendix E.

2024

Llama

Fortesting

BBH

BOOLQ

Dubey

2020

Ultramem

PKM

MoE

MOE

COVERTECTION

Fedus

2022

Ultramem

Seeapendix C

D

Su

2024

151m

m

.

b

6.5b

2

1.6B

0.01

Fedus

2022

Jiang et al

2024

Ultramem

= 0.001

= 0.15

E

Xiong et al

5.2

					1 3					
	11									
	Ultramem									
12				6.5B				1.6B		
			1							
Model	Param (B)	FLOPs (G)	Val. loss#	GPQA "	TriviaQA "	BBH cot "	Hella Swag "	Wino Grande "	DROP "	Avg "
Dense-151M	0.15	0.30	2.96	19.98	12.67	22.57	35.07	52.49	13.60	26.06
PKM-151M-x12	2.04	0.35	2.76	17.30	24.66	23.14	42.25	51.38	13.10	28.64
MoE-151M-2in32	2.04	0.35	2.63	17.30	33.27	23.24	48.44	55.96	18.57	33.20
UltraMem-151M-x12	2.03	0.35	2.67	19.42	28.97	22.65	43.96	50.83	14.08	29.99
Dense-680M	0.68	1.36	2.64	21.09	27.16	24.65	48.83	54.93	22.97	33.27
PKM-680M-x12	8.95	1.50	2.46	20.65	46.31	26.97	57.32	61.72	25.20	39.70
MoE-680M-2in33	8.95	1.50	2.39	20.54	34.19	26.63	62.71	59.98	26.54	38.43
UltraMem-680M-x12	8.93	1.49	2.37	21.99	55.17	26.62	64.15	60.54	25.14	42.27
Dense-1.6B	1.61	3.21	2.49	21.76	39.65	26.41	58.6	61.72	22.63	38.46
PKM-1.6B-x12	21.13	3.48	2.34	22.99	48.92	28.98	65.45	63.93	27.55	42.97
MoE-1.6B-2in34	21.36	3.52	2.30	21.32	59.56	29.46	67.34	63.93	28.81	45.07
UltraMem-1.6B-x12	21.41	3.50	2.24	24.66	66.38	30.63	71.52	66.38	29.99	48.26
Dense-6.5B	6.44	12.88	2.30	19.98	57.28	31.14	69.73	65.9	33.12	46.19

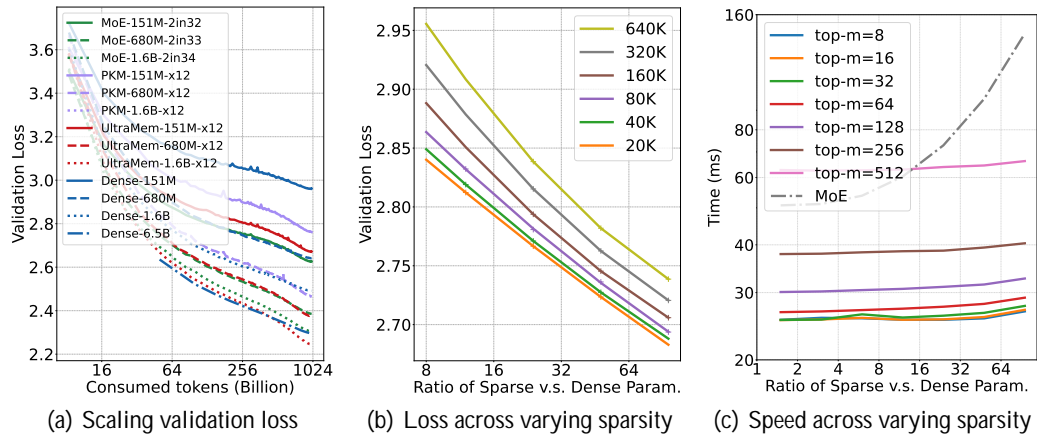
2

Allresults

3

7

all



6 a C
20 b 151m
20,000
c 512
1 Ultramem Moewith . B
2048 / Cachelength
MOE

5.3 M
LLM MOE Ultramem Weensparsity
Thechanges TOP-M 6 b

80K TheDefault
6 c TOP-M
Ultramem MOE
1 b MoE

5.4
151m ThePKM MLP
1.2E - 500B LR
2 C
1 6
2 Ultramem top-m

	Train Loss #	Valid. Loss #	Dense Param.(M)	Sparse Param.(G)	FLOPs (M)
PKM-151M-x10	2.604	2.828	173.01	1.534	346.06
+rm softmax	2.570 -0.034	2.822 -0.006	173.01	1.534	346.06
+half vdim+proj	2.556 -0.014	2.800 -0.022	178.47	1.529	356.98
+share query	2.560 +0.004	2.803 +0.003	173.46	1.529	346.96
+split big mem&skip	2.554 -0.006	2.788 -0.015	161.64	1.536	323.32
+query/key LN	2.553 -0.001	2.789 +0.001	161.64	1.536	323.54
+IVE	2.544 -0.009	2.772 -0.017	172.37	1.536	344.98
+TDQKR	2.538 -0.006	2.764 -0.008	172.37	1.536	344.98
+MCS	2.521 -0.017	2.761 -0.003	172.37	1.536	344.98
+improved init	2.518 -0.003	2.758 -0.003	172.37	1.536	344.98
+value lr decay	2.494 -0.024	2.736 -0.022	172.37	1.536	344.98
+query conv	2.493 -0.001	2.736 -0.000	172.38	1.536	345.02
Total Diff	-0.111	-0.092	-0.64	+0.002	-1.04

3 - r = 2

4 h = 2

5 e = 4

6 LR

10. b c

10. c

8

3 IVE TDQKR MC

3 IVE E EVER

InModel E = 4 TDQKR MCS R HDO

R = 2 H H = 2 H = 2

3 IVE TDQKR MC

	IVE				TDQKR				MCS			
	Baseline	E=4	E=9	E=16	Baseline	r=2	r=3	r=4	Baseline	h=2	h=4	h=8
Training loss [#]	2.553	-0.009	-0.016	-0.019	2.544	-0.006	-0.0065	-0.0063	2.538	-0.017	-0.017	-0.012
Validation loss [#]	2.789	-0.017	-0.025	-0.027	2.772	-0.008	-0.0084	-0.0082	2.764	-0.003	+0.001	+0.006
FLOPs(G)	323.54	+6.6%	+14.9%	+26.4%	344.98	+0.001%	+0.002%	+0.003%	344.98	+0.001%	+0.003%	+0.007%

10

ACKNOWLEDGMENTS

Pingshuo Ma Wenda Liu Ultramem
 Siyan Chen
 Fan Xia MOE

REFERENCES

- Hervé Abdi SVD 907 912
 44 2007
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.
- E. arXiv arXiv 1607.06450 2016
- Daniel Bershtatsky, Daria Cherniuk, Talgat Daulbaev, Aleksandr Mikhalev, and Ivan Oseledets. Lotr: Low tensor rank weight adaptation. *arXiv preprint arXiv:2402.01376*, 2024.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.
- Tom B Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.
- Zewen Chi Li Dong Shaohan Huang Damai Dai Shuming MA Barun Patra Saksham Singhal Payal Bajaj Xia Song Xian-Ling Mao
 35 34600-34613 2022
- Boolq / arXiv arXiv 1905.10044 2019
- Peter Clark Isaac Cowhey Oren Etzioni Tushar Khot Ashish Sabharwal Carissa Schoenick
 Oyvind Tafjord arc ai arXiv arXiv
 1803.05457 2018
- Redpajama llama 2023.URL [https://github.com/togethercomputer/](https://github.com/togethercomputer/RedPajama-Data)
 RedPajama-Data
- Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. Approximating two-layer feedforward networks for efficient transformers. *arXiv preprint arXiv:2310.10837*, 2023.
- Deepseekmoe arXiv arXiv: , 2024
- Dheeru Dua Yizhong Wang Pradeep Dasigi Gabriel Stanovsky Sameer Singh Matt
 Gardner Drop arXiv arXiv 1903.00161 2019
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 23 120 1-39 2022
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Xu Owen He. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*, 2024.

arXiv arXiv 2009.03300 2020

AG Howard Mobilenets ARXIV ARXIV
1704.04861 2017

Herve Jegou Matthijs Douze Cordelia Schmid
33 1 117–128 2010

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

arXiv arXiv: , 2017 Triviaqa

Gyuwan Kim Tae-Hwan Jung ArxivpreprintArxiv 2010.03881 2020

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024.

Guillaume Lample Alexandre Sablayrolles Marc' aurio Ranzato Ludovic Denoyer
Hervéjégou 2019 32 2019

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Osman Asif Malik Stephen Becker TensorSketch Tucker
2018 3 31

Deepak Narayanan Mohammad Shoeybi Jared Casper Patrick Legresley Mostofa
Patwary Vijaykorthikanti Dmitri Vainbrand Prethvi Kashinkunti Julie Bernauer
Bryan Catanzaro Bryan Catanzaro Gputron-eggutron-
tersters tersters pp 1-15 2021

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Ra el Noam Shazeer Adam Roberts Katherine Lee Sharan Narang Michael Matena
YanqiZhou Wei Li Peter J Liu
21 140 1-67 2020

Samyam Rajbhandari Reza Yazdani Aminabadi Am-mar
Ahmad Awan Je Rasley Yuxiong He Deepspeed-moe
18332–18346 PMLR 2022

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

17566 2021 34 17555–

Winogrande
winograd ACM 64(9): 99–106, 2021

- arXiv: , 2015
- arXiv
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- Kai Shen Junliang Guo Xu Tan Siliang Tang Rui Wang Jiang Bian Relu Softmax
Arxiv ARXIV 2302.06461 2023
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-Lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Mirac Suzgun Nathan Scales Nathanael Schärli Sebastian Gehrmann Yi Tay Hyung Won Chung Aakanksha Chowdhery Quoc V Le Ed H Chi Denny Zhou
arXiv arXiv: , 2022
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Transformer 10524–10533 PMLR 2020
- arXiv arXiv: , 2019 Yejin Choi Hellaswag
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.
- Quoc V Le James Laudon
35 7103–7114 2022

number, $1 \rightarrow$, We

Top-M

TOP-M

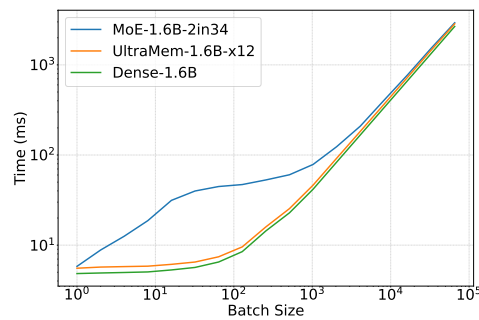
$\frac{1}{e} y$

$\frac{1}{dkto}$

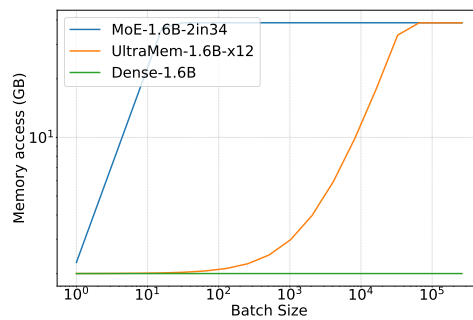
$\frac{1}{\overline{D_k}}$

B

7 MOE 131,072 Ultramem MOE



(a) Inference time



(b) Memory access

7 MOE Ultramem
MOE Ultramem
1
2048

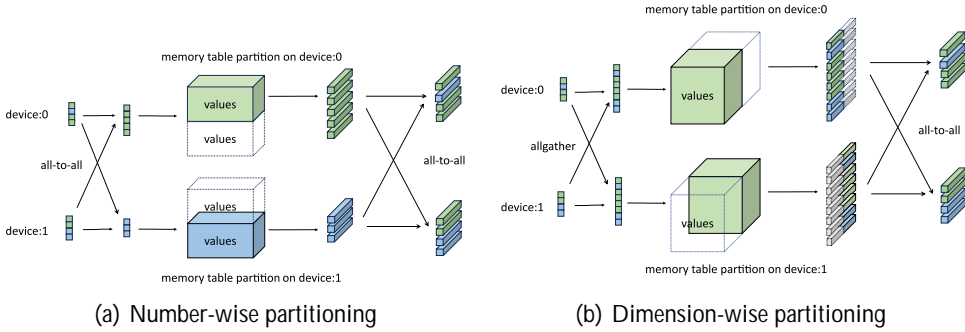
thex y /

C

2021 3D Megatron Shoeybi 2019; Narayanan

A single GPU
DATA

D Number



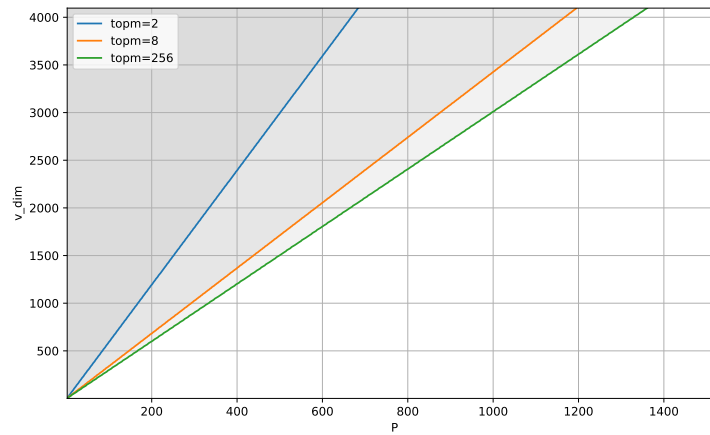
8

TOP-M

fordimension-wise

d

8



$$9 \quad p \quad v \quad / \quad 1 \quad / \quad 1$$

P

- All-to-all transmission of indices: $\text{sizeof}(\text{int}) \quad bs \quad \text{topm} \quad (P - 1) = P$
- $\text{sizeof}(\text{bfloat}) \times bs \times \text{topm} \times t \text{ topm} \times v \text{ dim} \times p - 1 / p$
- AllGather indices: $\text{sizeof}(\text{int}) \quad bs \quad \text{topm} \quad (P - 1)$
- AllGather scores: $\text{sizeof}(\text{bfloat16}) \quad bs \quad \text{topm} \quad (P - 1)$
- $b \text{ oat} \times vs \times v \text{ dim} \times p - 1 / p$

$$v \text{ dim} \quad BS \quad 9$$

$$P \quad V \text{ DIM}$$

E

4
 LR" Val-ues e-
 . E- e- 1.2e- vEN- 1.2E-
 151m m . . b 6.5b
 2020 Ultramem PKM
 Forultramem- m 3
 5/6 8/9 11 3 5 laylayer
 3 layerer 5
 Ultramem- m 3 7/8 12/13
 17/18 22 Ultramem- . b 3
 7/8 12/13 17/18 22/23 27/28 32
 forpkm- m 6 6
 PKM- M 12 12
 PKM- . B 16:16 Ultramem Moe
 6
 PA- -
 5 Ultra-MEM

Configuration Key	Value
Weight decay	0.1
1	0.9
2	0.95
LR	6e-4/2.5e-4/2e-4/1.2e-4
LR end ratio	0.1
LR schedule	cosine
LR warmup ratio	0.01
Dropout	0.1
Batch size	2048
Sequence length	2048
Training step	238418

4

Configuration Key	Value
Tucker rank r	2
Multi-core scoring h	2
Virtual memory expansion E	4
Aux loss weight	0.001
Aux loss margin	0.15

5 Ultramem

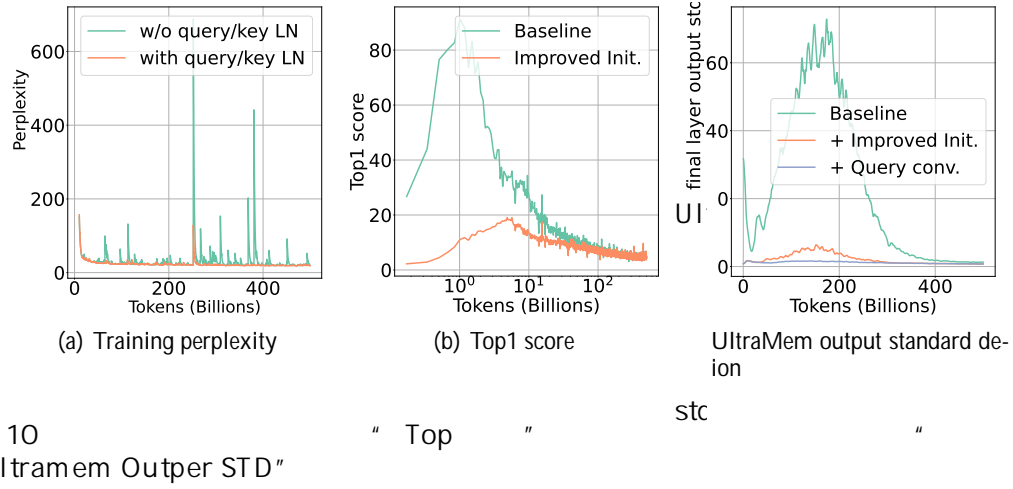
10

1. Knowledge: Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), TriviaQA (Joshi et al., 2017), Graduate-Level Google-Proof Q&A Benchmark (GPQA) (Rein et al., 2023), AI2 Reasoning Challenge (ARC) (Clark et al., 2018).
2. Reasoning: BIG-Bench Hard (BBH) (Suzgun et al., 2022), Boolean Questions (BoolQ) (Clark et al., 2019), HellaSwag (Hella) (Zellers et al., 2019), WinoGrande (Wino) (Sakaguchi et al., 2021).
3. Reading comprehension: Discrete Reasoning Over Paragraphs (DROP) (Dua et al., 2019).
4. Comprehensive ability: AGIEval (Zhong et al., 2023)

Model	Hidden Dim	Inner Dim	Attn Head	Layer	Top-m	Expert	Kdim	Knum	Mem Layer	Param (B)	FLOPs (G)
Dense-151M	1024	4096	16	12	-	-	-	-	-	0.15	0.30
Dense-680M	1536	6144	16	24	-	-	-	-	-	0.68	1.36
Dense-1.6B	2048	8192	16	32	-	-	-	-	-	1.61	3.21
Dense-6.5B	4096	16384	32	32	-	-	-	-	-	6.44	12.88
MoE-151M-2in32	1024	2528	16	12	2	32	-	-	-	2.04	0.35
MoE-680M-2in33	1536	3584	16	24	2	33	-	-	-	8.95	1.50
MoE-1.6B-2in34	2048	4672	16	32	2	34	-	-	-	21.36	3.52
PKM-151M-x12	1024	4096	16	12	16x6	-	512	1347	1	2.04	0.35
PKM-680M-x12	1536	6144	16	24	35x8	-	768	2308	1	8.95	1.50
PKM-1.6B-x12	2048	8192	16	32	42x12	-	896	1792	1	21.44	3.52
UltraMem-151M-x10	2048	8192	16	32	42x2	-	256	1024	3	1.71	0.35
UltraMem-151M-x12	1024	4096	16	12	16x2	-	256	1100	3	2.03	0.35
UltraMem-680M-x12	1536	6144	16	24	35x2	-	384	1632	4	8.93	1.49
UltraMem-1.6B-x12	2048	8192	16	32	42x2	-	448	1792	6	21.41	3.50

6 TOP-M MOE PKM Ultramem ValueNumber
 Times KDIM PKM knum knum 2

F

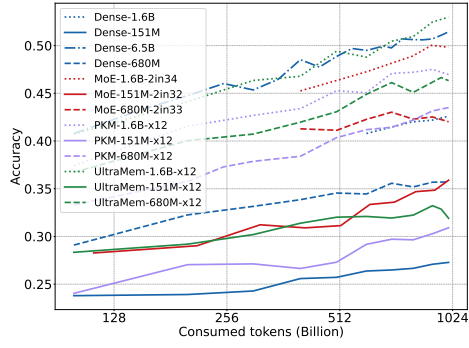


7

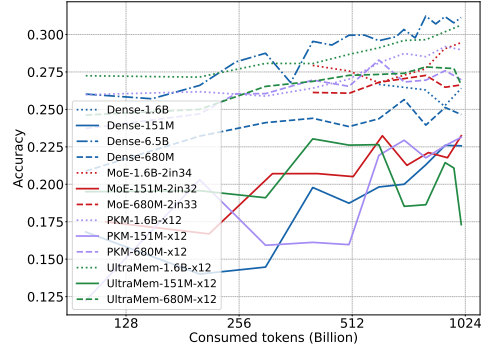
Model	Param	FLOPs	ARC-C	GPQA	Trivia	MMLU	BBH	BoolQ	Hella	Wino	AGI	DROP	Avg
Model	(B)	(G)			QA		cot		Swag	Grande	Eval		
Dense-151M	0.15	0.30	25.60	19.98	12.67	26.50	22.57	50.15	35.07	52.49	9.03	13.60	26.77
PKM-151M-x12	2.04	0.35	25.94	17.30	24.66	25.69	23.14	53.48	42.25	51.38	9.65	13.10	28.66
MoE-151M-2in32	2.04	0.35	26.96	17.30	33.27	26.58	23.24	55.96	48.44	55.96	9.34	18.57	31.56
UltraMem-151M-x12	2.03	0.35	25.68	19.42	28.97	25.62	22.65	47.74	43.96	50.83	10.00	14.08	28.89
Dense-680M	0.68	1.36	24.06	21.09	27.16	24.64	24.65	46.42	48.83	54.93	9.44	22.97	30.42
PKM-680M-x12	8.95	1.50	25.51	20.65	46.31	25.22	26.98	41.80	57.32	61.72	8.94	25.20	33.97
MoE-680M-2in33	8.95	1.50	25.17	20.54	34.19	24.38	26.63	43.70	62.71	59.98	7.39	26.54	33.13
UltraMem-680M-x12	8.93	1.49	23.72	21.99	55.17	24.97	26.62	48.20	64.15	60.54	8.26	25.14	35.88
Dense-1.6B	1.61	3.21	26.30	21.76	39.65	26.19	26.41	51.50	58.6	61.72	9.22	22.63	34.81
PKM-1.6B-x12	21.13	3.48	26.71	22.99	48.92	24.80	28.98	60.06	65.46	63.93	9.51	27.55	37.89
MoE-1.6B-2in34	21.36	3.52	25.43	21.32	59.56	26.18	29.46	42.78	67.34	63.93	6.63	28.81	37.14
UltraMem-1.6B-x12	21.41	3.50	25.94	24.66	66.38	24.67	30.63	59.8	71.52	66.38	8.77	29.99	40.88
Dense-6.5B	6.44	12.88	28.16	19.98	57.28	27.68	31.14	68.2	69.73	65.9	9.23	33.12	41.04

8

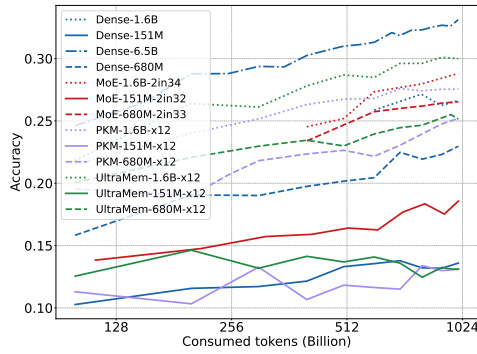
	Train Loss #	Valid. Loss #
PKM-151M-x10	2.604	2.828
+ rm softmax	-0.034	-0.006
+ half vdim+proj	-0.027	-0.02
+ share query	-0.003	-0.002
+ split big mem	-0.003	-0.005
+ query/key LN	-0.002	+0.003
+ IVE	-0.025	-0.023
+ TDQKR	-0.003	-0.007
+ TDQKR + MCS	-0.02	-0.009
+ value lr decay	-0.017	-0.007
+ query conv	-0.005	-0.001



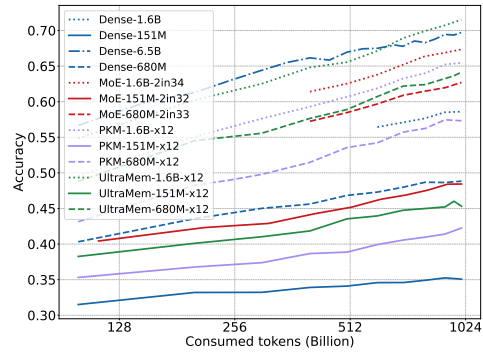
(a) Average accuracy



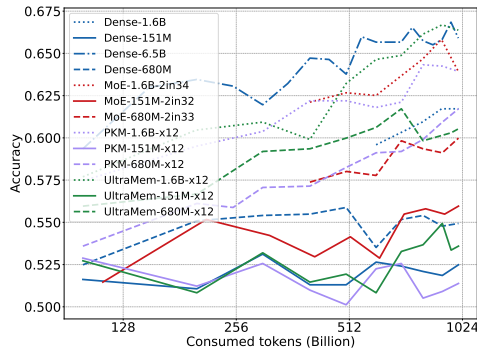
(b) BBH-cot-3shot accuracy



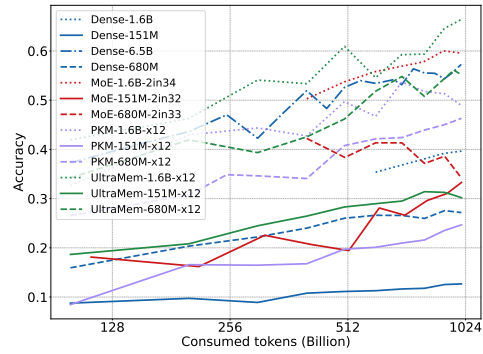
(c) DROP accuracy



(d) Hellaswag accuracy



(e) Winogrande 5shot accuracy



(f) TriviaQA 5shot accuracy