Beyond Algorithms: Navigating the Data Deluge in Al

刘群 LIU Qun

Huawei Noah's Ark Lab

ICLR 2024 EXPO Talk by Huawei

2024.05.08, Vienna, Austria





Pangu Large Models: Al4Industry

- Huawei regards AI as a huge and crutial oppotunity for the future of the company.
- Huawei has invested the full-stack AI technologies for AI, including NPU chips (Ascend), clusters (Atlas), AI frameworks (MindSpore), AI models (Pangu), and a broad spectrum of AI applications, especially for industries.





Beyond Algorithms: Navigating the Data Deluge in Al

Building large-scale AI models has become a massive systems engineering problem, far more than just an algorithm problem, which requires cooperations among scientists and engineers from multiple disciplines.



Data Management: acquisition, cleaning, labeling, and pre-processing





PanGu- Σ : dense-sparse architecture with heterogeneous computing





PanGu-Σ: Towards Trillion Parameter Language Model with Sparse Heterogeneous Computing, arXiv:2303.10845



CAME: confidence-guided adaptive memory efficient optimization



arXiv:2307.02047 ACL 2023 Outstanding Paper Awards



Efficient inference: quantization, compression and deployment

1~4x memory usage reduction, 100% throughput improving

Challenges in inference: Large number of parameters, slow inference,

high memory usage, high cost in end-to-end inference

- > Traditional quantization causes significant precision degradation for generative models
- High memory usage in inference: (1) Model parameters: 350 GB memory for a 175B model.
 2) KV cache: 576 GB for a 175B model with a 4 KB length.

Low-bit Weighting Algorithm: QuantGPT

Progress: (1) 4/8-bit quantization algorithm (2) Ascend affinity efficient dequantization operator (3) 2-4x memory reduction, 15-30% inference acceleration. Deployment of a 388 model on a single card.



arXiv:2203.10705v2 ACL 2022 Outstanding Paper Awards

KV cache Compression







Separate deployment, Dynamic batch

Progress: Separate deployment of full and incremental inference (8+8), 30-50% throughput improvement

Dynamic batch:

• Early exit for completed samples • New samples added in time Separate deploy for full and

incremental inference :

- Full inference: batch size=1 to reduce delay
- Incremental inference: Large Batch Size to improve throughput





Efficient Post-Training Pruning Method for LLMs



Advantages of LLM pruning

	Memor	y Access	Computation		
	Model Size	Throughput	Prefil	Decode	
Unstructured Pruning	+	1			
Structured Pruning	+	+	+	+	
Semi-structured Pruning	+	T	4	4	

Challenges in LLM Pruning: Channel Collapse



A new pruning metric

Method

2



• Relative Importance and Activation (RIA) : $\mathbf{RIA}_{ij} = \mathbf{RL}_{ij} \times (||\mathbf{X}_i||_0)^a = (\frac{||\mathbf{W}_{ij}||}{2} + \frac{||\mathbf{W}_{ij}||}{2}) \times (||\mathbf{X}_i||_0)^a$

$$\mathbf{IA}_{ij} = \mathbf{RI}_{ij} \times (||\mathbf{X}_i||_2)^a = (\frac{1}{\sum |\mathbf{W}_{*j}|} + \frac{1}{\sum |\mathbf{W}_{*i}|}) \times (||\mathbf{X}_i||_2)^a$$

jointly normalizes the weight in the input and output dimensions, together with activations

Effectively resolving the channel collapse issue

Semi-structured pruning (2:4/4:8 sparsity)

- Permute the channels equivalently to find a better 2:4/4:8 sparse pattern
- Fast permutation: 15.3s for a single linear layer, 100x faster than previous permutation method

Result

Unstructured Pruning

Table 1: Perplexity results on Wikitext2. We produce the one-shot Post-Training pruning methods with 50% unstructured sparsity on LLaMA, LLaMA2 and OPT models.

Method	LLaMA 7b	LLaMA 13b	LLaMA 30b	LLaMA 65b	LLaMA2 7b	LLaMA2 13b	LLaMA2 70b	OPT 1.3b	OPT 13b
Dense	5.68	5.09	4.77	3.56	5.47	4.88	3.32	14.62	10.13
Magnitude	17.28	20.22	7.54	5.90	16.02	6.83	5.36	1712	11561
Wanda	7.26	6.15	5.24	4.57	6.92	5.99	4.22	18.41	11.92
SparseGPT	7.24	6.20	5.32	4.57	6.99	6.10	4.25	27.00	11.18
RIA (Ours)	7.12	6.08	5.08	4.38	6.81	5.83	4.11	18.08	11.05

- SOTA performance under 50% sparsity for 7B-65B model
- No parameter update and all sizes of LLMs can be compressed within seconds

Semi-structured Pruning (2:4/4:8)

Table 5: LLaMA2-70B: Zero-Shot Performance of N:M constraint model comparing to the dense model. Bold values denote the best performance across all N:M constraint models. An asterisk ("*") signifies performance supposing that of the dense method.

Method	Hellaswag	BoolQ	ARC-C	MNLI	RTE	AVG	
Dense	64.77	83.70	54.44	45.81	67.87	63.32	i.
Wanda (2:4)	57.35	81.44	46.01	37.69	68.59*	58.22	0
Wanda (2:4+CP)	59.37	84.50*	48.55	43.09	66.43	60.39	
Wanda (4:8+CP)	60.86	82.73	49.94	40.15	67.87	60.51	
RIA (2:4)	57.13	82.78	46.76	37.39	69.31*	58.68	
RIA (2:4+CP)	58.48	85.14*	49.15	49.08*	68.95*	62.16	
RIA (4:8+CP)	60.44	83.58	50.43	48.69*	70.04*	62.64	

- Only ~1% average Acc drop
- The channel permutation semi-sparsity only takes 40 minutes (LLaMA 65B)

Zhang, Y., et.al, Plug-and-Play: An Efficient Post-Training Pruning Method for Large Language Models. ICLR 2024 Hall B #225, 4:30PM – 6:30 PM



IntactKV: An orthogonal solution to enhance guantized LLMs



Consistent improvement over existing methods on MMLU benchmark

Task Acc	MMLU (5 shot) average						
Vicuna Family	v1.5-7B	v1.5-13B	v1.3-7B	v1.3-13B	v1.3-33B		
FP16	49.84%	55.78%	47.12%	52.10%	59.30%		
RTN	44.62%	51.44%	39.33%	44.56%	53.18%		
GPTQ	43.99%	52.95%	40.12%	47.83%	55.84%		
OmniQuant	46.54%	52.86%	43.18%	47.92%	55.12%		
AWQ	46.45%	52.92%	43.08%	48.56%	56.09%		
+INTACTKV _[B]	46.87%	53.58%	44.67%	49.05%	56.91%		

Integratable with KV Cache Quantization



Other advantages:

- Orthogonal to existing quantization methods
- Plug-and-play: no extra training /inference

Liu, R., et.al, IntactKV: Improving Large Language Model Quantization by Keeping Pivot Tokens Intact, arXiv: 2403.01241, 2024



SELF: self-improving and self-evolving for LLMs

- Freely available high quality data is going to be exhausted in the near future.
- Instruct data for SFT and human preference data for RLHF are expensive.
- We introduce an innovative approach, SELF, which empowers LLMs to undergo continual self-evolution, thereby augments their inherent capabilities.



SELF: Self-Evolution with Language Feedback, arXiv:2310.00533v2



Gaining Wisdom from Setbacks: Aligning LLMs via Mistake Analysis



Gaining Wisdom from Setbacks: Aligning LLMs via Mistake Analysis, ICLR 2024, arXiv:2310.10477v1



RAG with Vector Database

Application Case: Telecom Knowledge Q&A

Telecom Knowledge QA

- Huge data size: >40M documents
- Complex format: lots of digits, tables, formula
- Complex questions: requiring precise knowledge to answer
- Digital sensitive

Question	What is the diameter of the fiber core in the optical module fiber of HiSecEngine USG6603F-C Tianguan product in international standard?	Query Understanding
Answer	There are international standard for the diameter of the fiber core in optical fibers: 62.5um and 50um for multimode, 9um for single-mode.	Query Embedding Query Keywords
Fine-tuned Pangu	The international standard of the diameter of the fiber core in the optical module fiber of HiSecEngine USG6603F-C Tianguan product is 125 micrometers.	Gauss Victor DP
+Gauss VectorDB	The international standard of the diameter of the fiber core in the optical module fiber of HiSecEngine USG6603F-C Tianguan product is 62.5um and 50um for multimode, and 9um for single-mode.	Answer Composition
Telecom Knowle	dge QA scores increased from 10.5 to 96, reaching the expert level.	Answer

The Pangu model has passed HCIE (Huawei Certified ICT Experts), surpassing more than 95% human experts.



Query

Pangu Agent for Operation Research Problem Solver

Application Case: Supply Chain Optimization

- Pangu Language Models: Problem Modeling
 - Dialog

X

42

- Complex reasoningCode generation
- Summarization

- OPTV Al Solver (Plug-ins):
 Problem Solving
 - Interpreting
 - Calculation
 - Optimization
 - Math problem solving

For level-1 problems of general operation research, we reaches a comparable level with GPT-4.

Training data	32,108	
Test data	289	
GPT-4	0.7048	
Pangu Production Model	0.6989	

The supply chain planner needs to make decisions on the allocation of key components (raw materials), SC (structural component) 1 and 2. There are three requirement codes, namely INV (inverter) A, B, and C. The requirements are 50 pieces, 100 pieces, and 150 pieces, respectively. The virtual material cost is 13, 2:1 and 18 respectively. The inventory of raw material SC 1 and 2 is 100 pieces and 90 pieces, respectively. SC 1 can be used to produce1NV A or B, and processing unit INV A or B requires two SC 1; Processing unit INV C requires one SC 2, and SC 2 can replace SC 1; but SC 1 cannot replace SC 2. May I ask how to make a decision on the production quantity of INVS A, B, and C, as well as the allocation of raw material SC 1 and 2, in order to maximize the total virtual material cost of the requirements met7 Build a mathematical planning model for this problem.



第△Ⅰ式解别





Pangu Al4Industries: 1000+AI Projects in 10+Industries implemented

Industrial Simulation: COMAC "Oriental Wing Wind"

Aerodynamic AI simulation of large passenger aircraft wing.

The average error of flow field prediction is as low as 0.001.

The single simulation speed is improved by 1000 times.

SAIL Award - WAIC2023 Top Award

EDA -

Large Language Model Code Generation.

Test sample generation coverage reaches 99.5%.

E2E efficiency of test R&D improved by >3x

Meteorology: Publication in "Nature"

24-hour global weather forecast in 1.4 seconds.

10,000 times faster than traditional methods.

The error is reduced by more than 12%.

Computing power consumption reduced by 600,000 times.

Extreme weather forecasts increased by 25%

Drug Discovery

Significantly shorten the drug development cycle. New broad-spectrum antibiotics were discovered in Xijiao University Affiliated Hospital within a month.

Industrial manufacturing: automatic production scheduling

The time for production line allocation schedule is

reduced from several hours to 1 minute.

- Government Affairs

Automatic scheduling of thousands of

back-end applications.

Quick realization of various services in cities.



Thank you!

把数字世界带入每个人、每个家庭、 每个组织,构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

Copyright©2018 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

