



Published as a conference paper at ICLR 2025

TRAINING-FREE ACTIVATION SPARSITY IN LARGE LANGUAGE MODELS

无训练激活稀疏加速大模型推理

汇报人:xx

时间:2025.12.03





目录

CONTENTS

01 背景与动机

02 方法总览

03 系统实现

04 实验评估

05 总结与展望





01

背景与动机





LLM推理的 Memory-Bound挑战

在小batch场景下，大模型推理主要受限于**显存带宽**而非算力，权重搬运成为瓶颈。

传统加速手段

如INT8/INT4量化、权重剪枝，仅对“模型参数”动刀，未能挖掘“中间激活”的冗余潜力。

新的加速思路

利用激活本身的结构特性，在推理时减少内存访问量，开辟一条**无需重训、即插即用**的加速新路径。



激活稀疏：被忽视的加速富矿



传统ReLU网络

天然95%稀疏，易于加速。



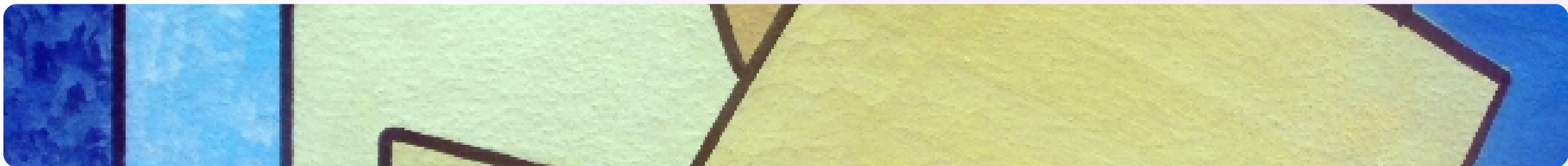
现代LLM (SwiGLU)

连续分布，天然稀疏性消失。



TEAL方案

无训练稀疏，实现1.5-1.8×加速。



02

方法总览

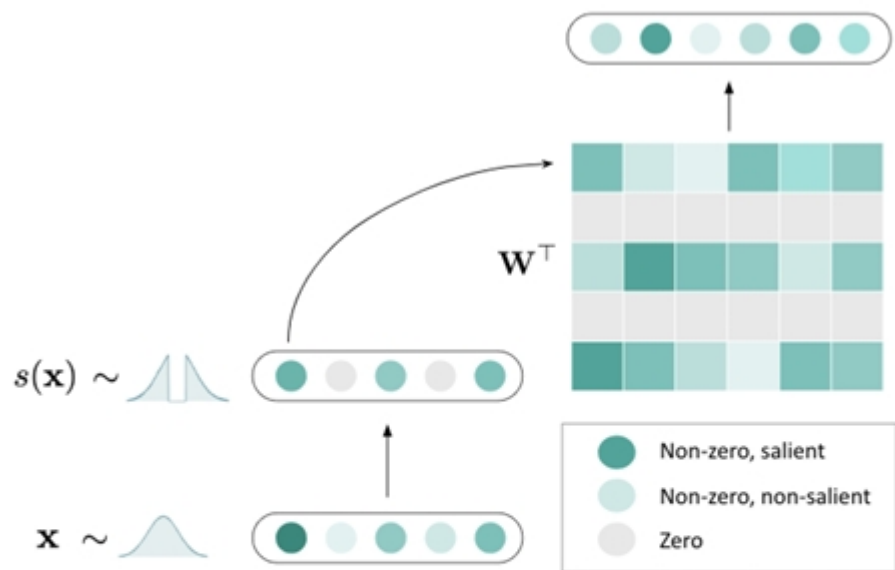


Figure 1: **Overview of TEAL.** During decoding, TEAL thresholds low-magnitude activation entries to zero, which obviates the need to move the associated weight channels onto the registers, thus enabling wall-clock speed-ups.

TEAL核心：无训练全局稀疏

- 🚫 **不改结构：**保持原有模型架构不变。
- ▶ **不再训练：**无需额外微调或预训练。
- 📊 **利用统计：**基于激活值在零附近密集分布的特性。
- ✂ **全局裁剪：**直接裁掉绝对值最小的激活，达成约 50% 全局稀疏。



分位数阈值裁剪机制

1. 统计分布

2. 确定阈值

3. 在线裁剪

离线采样，计算每层激活的绝对值分布

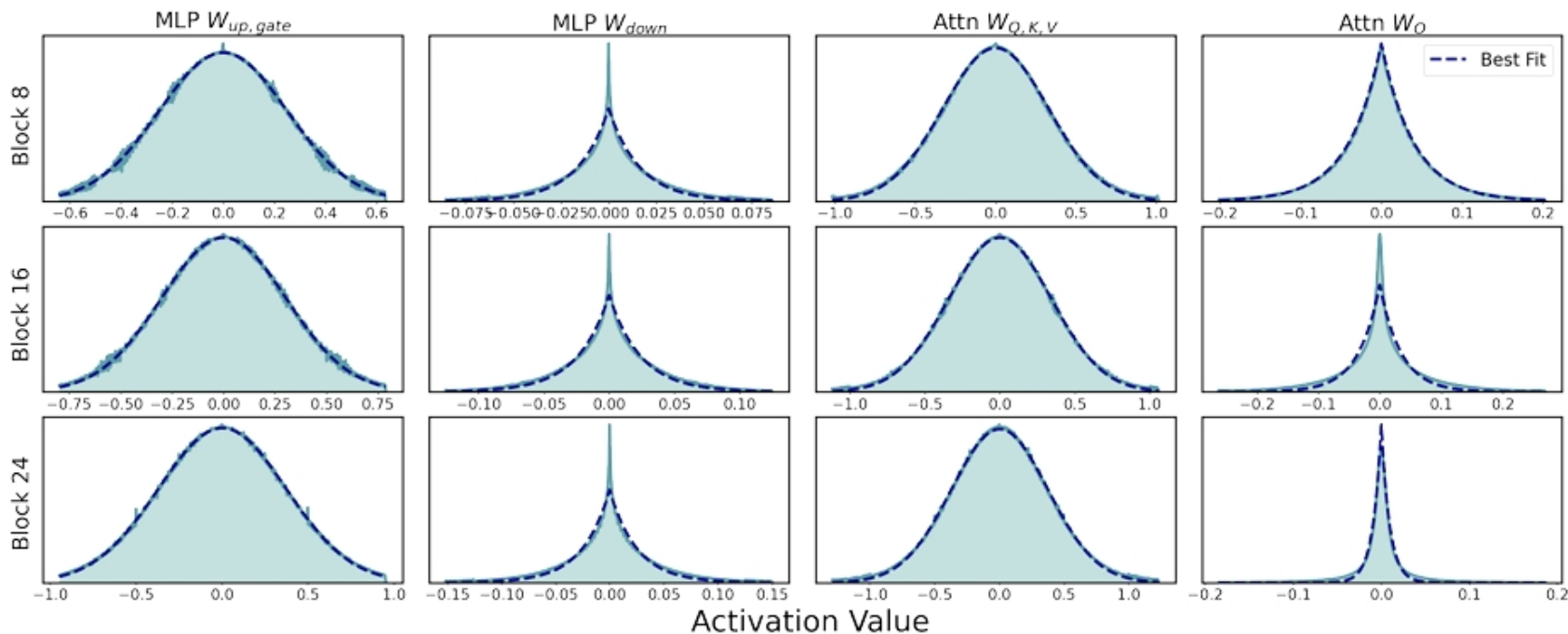


Figure 2: Activation distributions of Llama-3-8B's four hidden states at Blocks 8, 16, and 24. The activations preceding the Attention and MLP blocks typically exhibit Gaussian-like shapes, while intermediate states within these blocks exhibit Laplacian-like shapes. The best-fit Gaussian/Laplace distributions are overlayed in blue.



Block-wise 贪心分配稀疏度

以Transformer Block为单位，智能分配稀疏预算，保护关键矩阵。

Attention 矩阵

W_q, W_k, W_v, W_o

MLP 矩阵

$W_{gate}, W_{up}, W_{down}$

贪心算法 (Algorithm 1)

逐次将 5% 的稀疏增量分配给带来最小输出误差的矩阵，直至总稀疏度达标。结果揭示 **Wdown** 最易稀疏，**Wup** 最难剪。

03

系统实现



Sparse GEMV Kernel设计

仅置零无法省带宽，TEAL定制稀疏GEMV，让零值列真正跳过显存搬运，把稀疏转化为实测加速。



列存储 (Column-Major): 权重按列存储，方便按通道选择性加载。



操作融合: 掩码生成与矩阵乘融合，减少中间内存读写。



缓存优化: 调整缓存策略，让激活尽量留在L2 cache。

04

实验评估



实验设置：模型与评估



模型

Llama-2 (7B, 13B, 70B)

Llama-3 (8B, 70B)

Mistral-7B



语言建模

WikiText 数据集

困惑度 (Perplexity)



下游任务

MMLU, ARC-C, GSM8K等6项基准

平均准确率

对比方案：原始Dense模型 vs. 现有Training-Free方法 **CATS**



精度与稀疏度的权衡

Table 1: Perplexity results. Results between Llama-3 and Llama-2/Mistral are not directly comparable due to differing vocabulary sizes.

Method / Model	LLaMA-3		LLaMA-2			Mistral
	8B	70B	7B	13B	70B	7B
Baseline (0%)	5.87	2.93	5.07	4.50	3.12	4.92
CATS 25%	6.78	3.64	5.52	4.99	3.42	5.87
TEAL 25%	5.94	3.02	5.09	4.51	3.13	5.01
CATS 40%	$7.6 \cdot 10^4$	96.97	43.8	53.9	171	$2.8 \cdot 10^4$
TEAL 40%	6.21	3.52	5.22	4.60	3.25	5.13
TEAL 50%	6.67	4.30	5.43	4.76	3.50	5.31
TEAL 65%	9.06	6.29	6.62	5.50	4.28	6.23



精度与稀疏度的权衡

Table 2: Downstream task evaluation results. Reported results are averaged over six tasks. See Appendix [A.2](#) for fine-grained results. We omit CATS 40% as it is degenerate.

Method / Model	LLaMA-3		LLaMA-2			Mistral
	8B	70B	7B	13B	70B	7B
Baseline (0%)	68.07	80.41	56.50	62.01	72.65	66.96
CATS 25%	64.15	79.25	54.60	60.48	71.93	64.25
TEAL 25%	67.73	80.22	56.42	62.21	72.67	66.63
TEAL 40%	66.21	79.29	55.45	61.27	72.57	65.46
TEAL 50%	63.42	78.26	54.26	60.41	72.02	64.16
TEAL 65%	52.59	73.07	48.16	55.71	69.30	58.93

TEAL vs CATS

在相同稀疏度下，TEAL的困惑度和准确率普遍更优。其全局均衡稀疏策略，避免了局部过度剪枝带来的信息瓶颈。



解码加速与量化叠加

解码速度 (单batch)

在A6000上, 50%稀疏度可达 **1.5-1.8×** 加速。

Table 3: Single-batch end-to-end inference speed results, measured in tokens per second. We exclude Mistral-7B and Llama-2-70B as they are architecturally similar to Llama-3-8B and 70B. We utilize tensor parallelism for Llama-3-70B: TP2 for A100, and TP4 for A6000.

GPU	Sparsity	LLaMA-3		LLaMA-2	
		8B	70B	7B	13B
A6000	Baseline	45.32 (1.00×)	15.93 (1.00×)	50.54 (1.00×)	26.43 (1.00×)
	0%	44.49 (0.98×)	15.57 (0.98×)	50.06 (0.99×)	26.25 (0.99×)
	25%	55.38 (1.22×)	18.93 (1.19×)	64.54 (1.28×)	33.67 (1.27×)
	40%	64.15 (1.42×)	20.86 (1.31×)	77.30 (1.53 ×)	40.20 (1.52×)
	50%	73.94 (1.63×)	23.77 (1.49×)	89.91 (1.78×)	47.60 (1.80 ×)
A100	Baseline	100.79 (1.00×)	21.85 (1.00×)	110.15 (1.00×)	61.01 (1.00×)
	0%	92.13 (0.91×)	20.32 (0.93×)	100.97 (0.92×)	56.33 (0.92×)
	25%	112.11 (1.11×)	25.18 (1.15×)	126.14 (1.15×)	70.66 (1.16×)
	40%	126.24 (1.25×)	28.78 (1.32×)	143.85 (1.31×)	81.90 (1.34×)
	50%	134.29 (1.33×)	29.99 (1.37×)	154.02 (1.40×)	88.38 (1.45×)



解码加速与量化叠加

与权重量化兼容

与8bit RTN、4bit AWQ等量化方法叠加，误差相互独立，可联合使用，实现成本极限压缩。

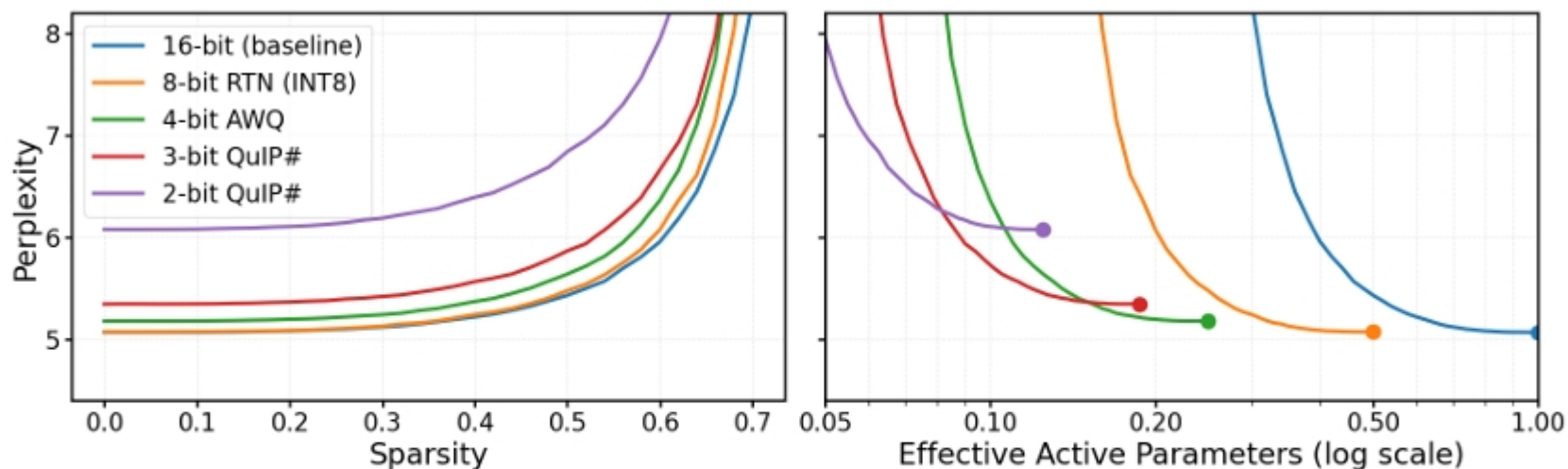


Figure 4: Perplexity vs. sparsity for Llama-2-7B quantized to various bitwidths on WikiText. Left: Performance over sparsity levels. Right: Performance normalized by bitwidth.

05

总结与展望



TEAL的价值与局限

核心价值

- **完全无训练**：零额外成本，即插即用。
- **全模型稀疏**：40-50%稀疏度下保持高精度。
- **显著加速**：单batch解码达1.5-1.8×提速。
- **兼容量化**：与权重量化正交，效果叠加。

主要局限

- **大batch效果减弱**：样本间激活模式差异大，难维持高稀疏。
- **Prefill阶段受限**：开头token的KV不能过度稀疏，否则损伤大。
- **适用场景特定**：主要利好小batch/单batch推理场景。



总结与启示

TEAL证明“中间结果”同样可被**聪明偷懒**，通过统计裁剪与kernel协同，把理论稀疏变为实际速度提升。



新范式

提供“只改推理栈、不动训练流程”的轻量级优化路线。



互补性

与主流的“只优化参数”方法形成完美互补。



实用性

对私有部署、边缘计算等场景有巨大吸引力。



感谢观看

汇报人:xx

时间:2025.12.03

