神经网络低比特量化

郭晋阳

北京航空航天大学人工智能学院

复杂关键软件环境全国重点实验室

北京航空航天大學

BEIHANG UNIVERSITY

复杂关键软件环境全国重点实验室

State Key Laboratory of Complex & Critical Software Environment

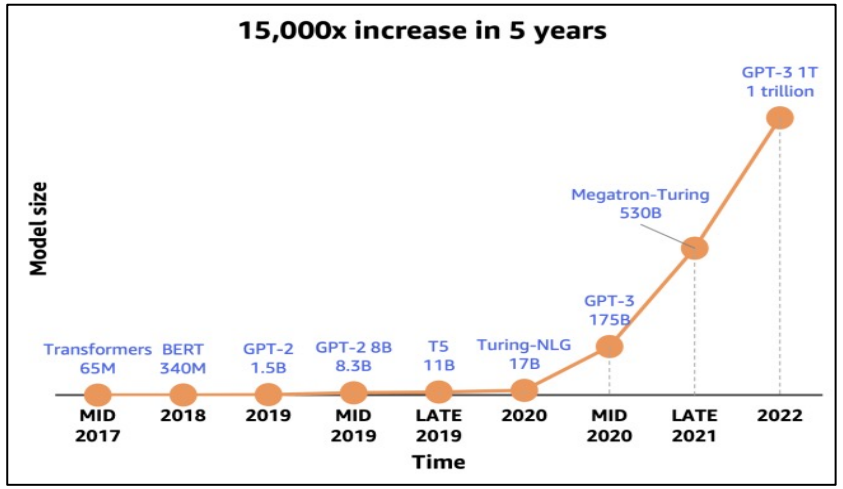
目录

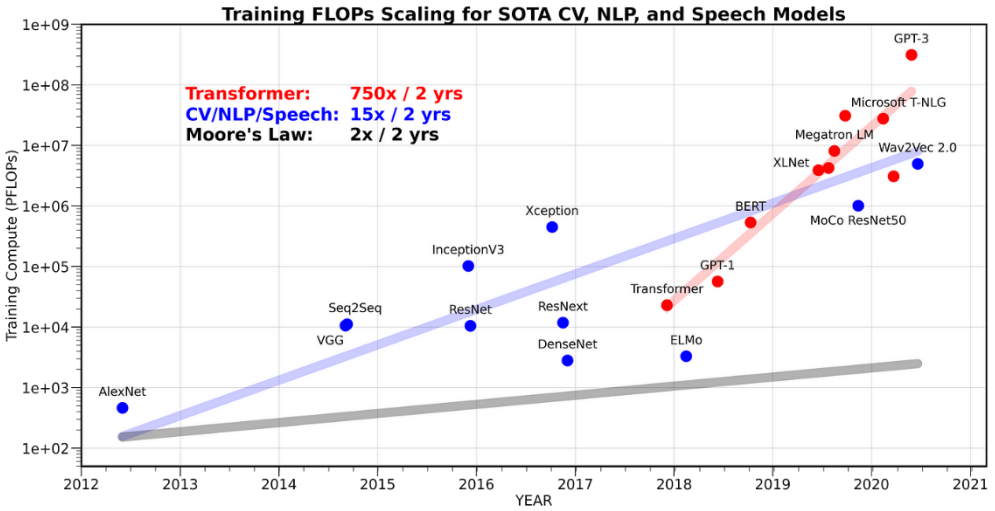
背景与挑战

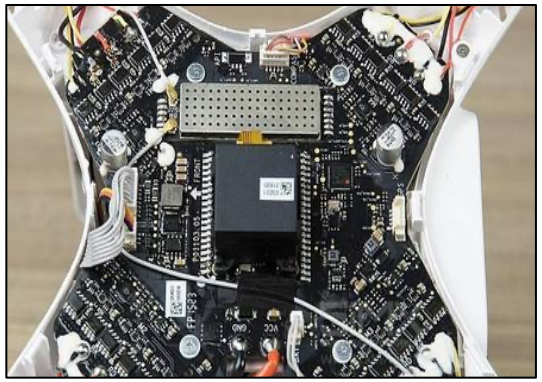
训练数据 **多**
模型参数 **多**

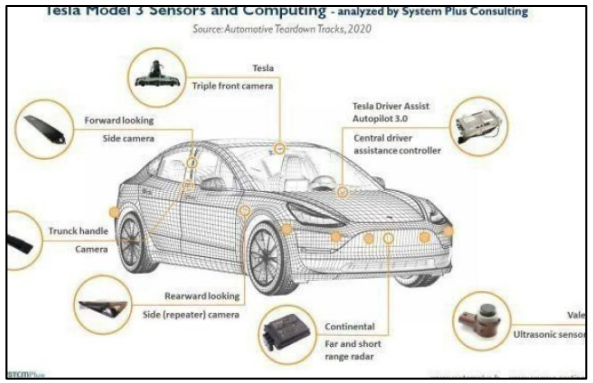
原型与部署使用存在巨大鸿沟

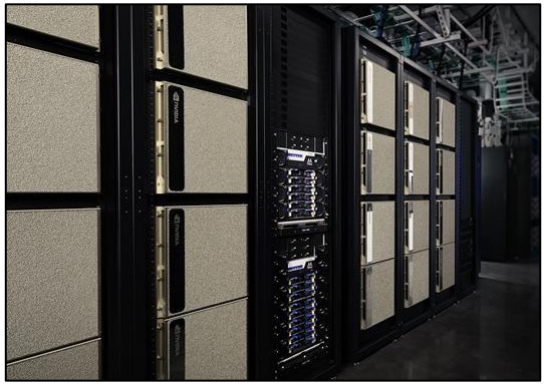
部署**场景多样**
片上**资源受限**



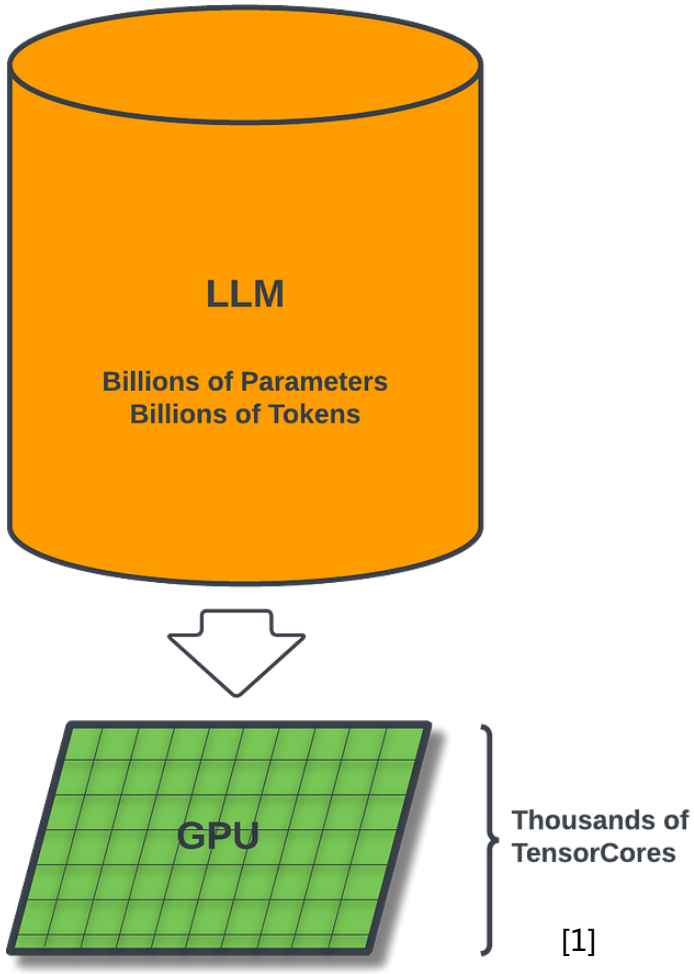








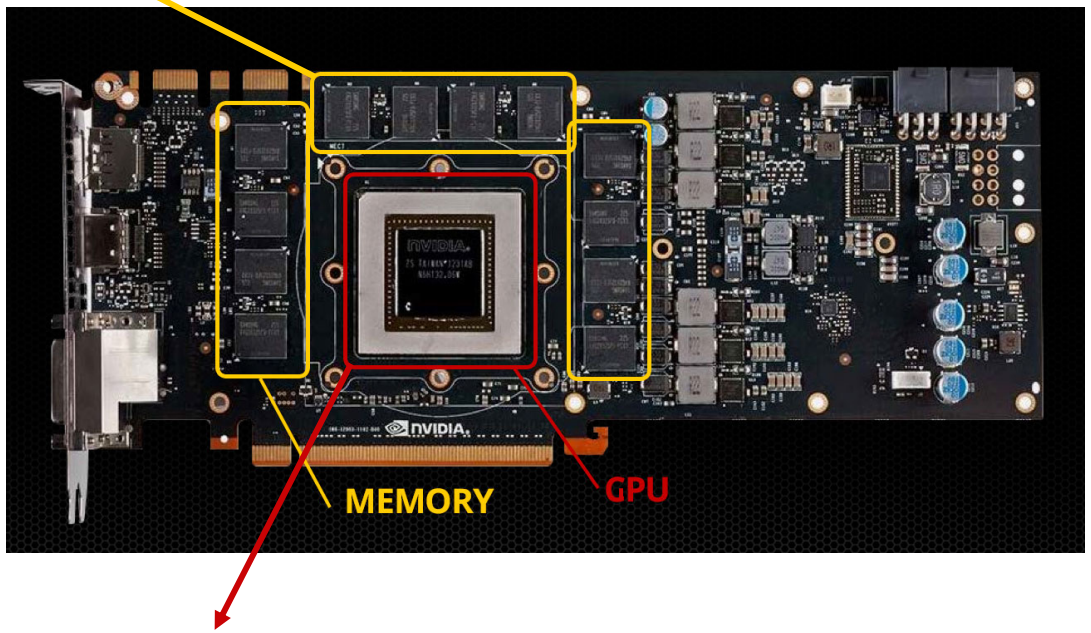
背景与挑战



[1]

显存占用: **196.52 GB** (FP16/BF16, w/o datasets) *vs.* 141GB (Nvidia H200)

LLaMA-13B
hidden_size: 5120
model_max_length: 4096
intermediate_size: 13696
num_attention_heads: 40
num_hidden_layers: 40
batch_size: 1



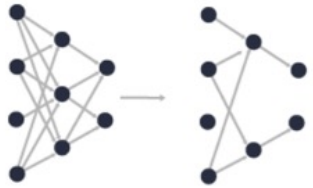
训练时间: **135,168 GPU 小时** (NVIDIA A100-80G) [2]

[1] Performance bottlenecks in deploying LLMs—a primer for ML researchers. 2023
[2] LLaMA: Open and Efficient Foundation Language Models. 2023

神经网络压缩

- **剪枝和参数共享**

  移除多余参数



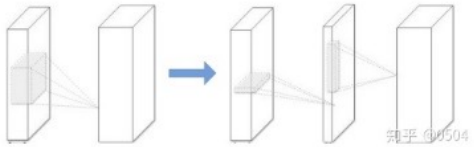
- **结构化卷积滤波器**

  替换过参数化的参数

$$\mathbf{R} = \mathrm{circ}(\mathbf{r}) := \begin{bmatrix} r_0 & r_{d-1} & \cdots & r_2 & r_1 \\ r_1 & r_0 & r_{d-1} & & r_2 \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{d-2} & & \ddots & \ddots & r_{d-1} \\ r_{d-1} & r_{d-2} & \cdots & r_1 & r_0 \end{bmatrix}.$$

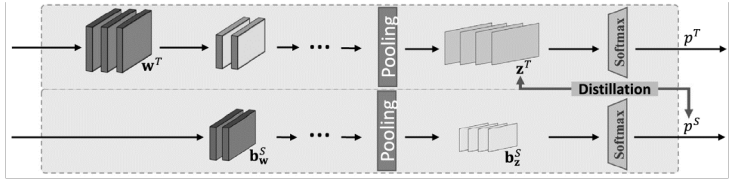
- **低秩分解**

  压缩数据表达



- **知识蒸馏**

  从大模型传递知识到小模型中



- **量化**

  低比特（有限位宽）表达
  例如 1-bit: 二值化，4/8-bit: INT4/8)

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit　8-bit　1-bit

Arch	1-bit	4-bit	8-bit
NvGPU-tn	128x	32x	16x
Arm v8.1	5.33x	N/A	1~2x
Arm v8.2	10.67x	N/A	4x
X86 avx2	5.33x	N/A	2~4x
X86 avx512	16x	N/A	4x

常见硬件架构低比特计算的峰值性能对比（相比fp32）

低比特量化

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit　　8-bit　　1-bit

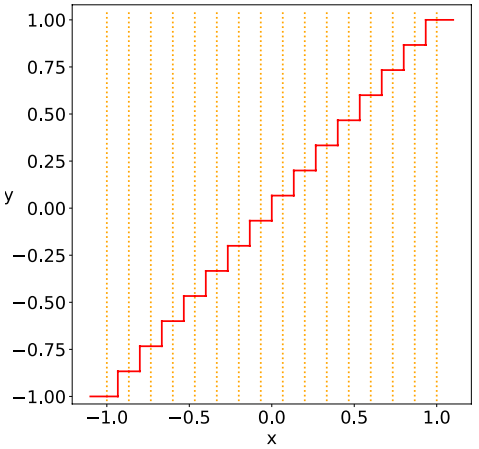
一、线性量化（2-8bit）

量化函数：

$$x_{int} = round(\frac{x}{\Delta}) + z$$

$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

反量化函数：$x_{float} = (x_Q - z)\Delta$



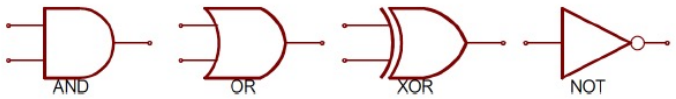
线性量化

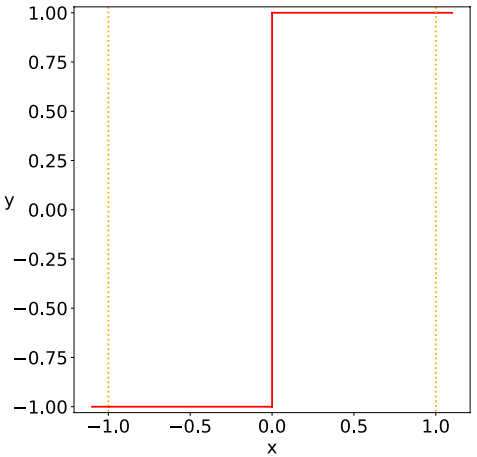
二、二值量化（1-bit）

量化函数：

$$Q_B(x) = \text{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

按位运算：

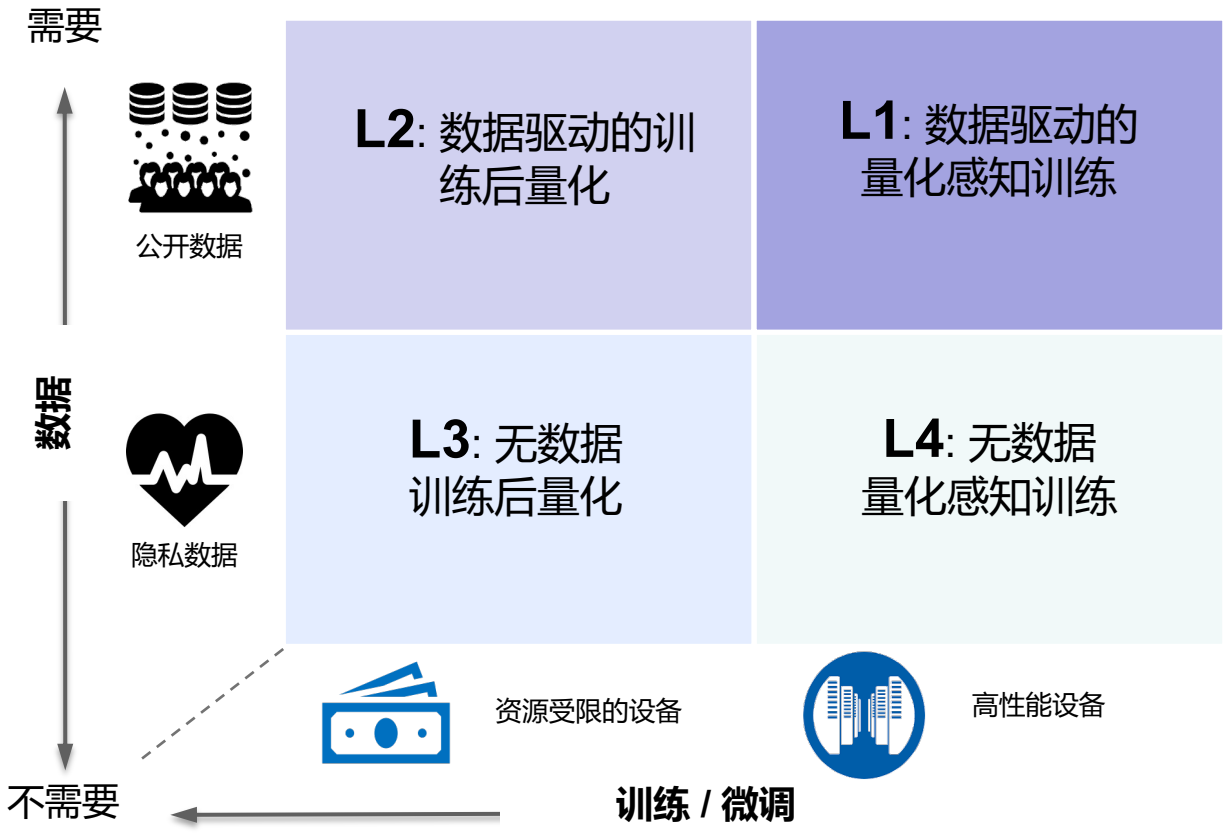




二值量化

低比特量化

- **挑战：无数据、低算力等限制场景下的模型量化生产和优化**



Post Training Quantization (PTQ)

Quantization Aware Training (QAT)

目录

相关工作

线性量化（2-8bit）：实用化低比特量化，面向更广泛场景

L2: 数据驱动的训练后量化	L1: 数据驱动的量化感知训练
L3: 无数据训练后量化	L4: 无数据量化感知训练

L2:

TFMQ-DM [Huang et al., BUAA. CVPR 2024]
Reg-PTQ [Ding et al., BUAA. CVPR 2024]
PTQ4SAM [Lv et al., BUAA. CVPR 2024]
…

4-bit PTQ [Banner et. al. Intel. 2019]
BRECQ [Li et. al. UESTC. 2020]

L1:

DSQ [Gong et al., BUAA. ICCV 2019]
INT8 training [Feng et al., BUAA. CVPR, 2020]

Integer-only [Jacob, et. al. Google. 2018]
LSQ [Esser. IBM. 2020]

L3:

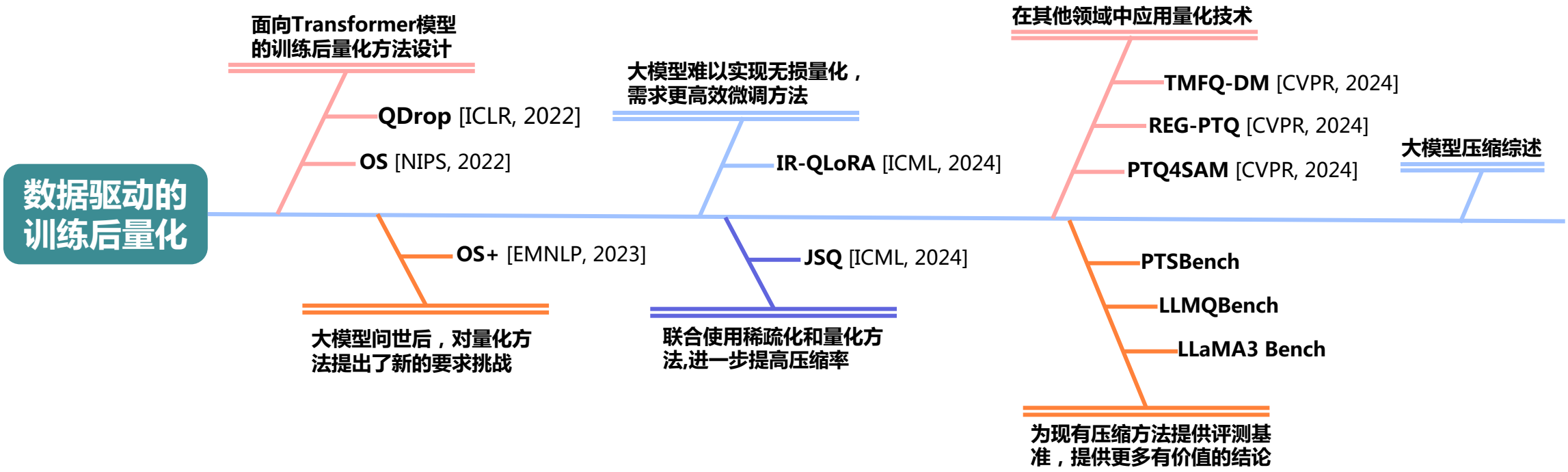
DSG [Zhang et al., BUAA. CVPR 2021]

ZeroQ [Cai et. al. PKU, UC Berkeley. 2020]
DFQ [Nagel et. al. Qualcomm. 2019]

L4:

DSG [Qin et al., BUAA. IEEE TPAMI 2024]

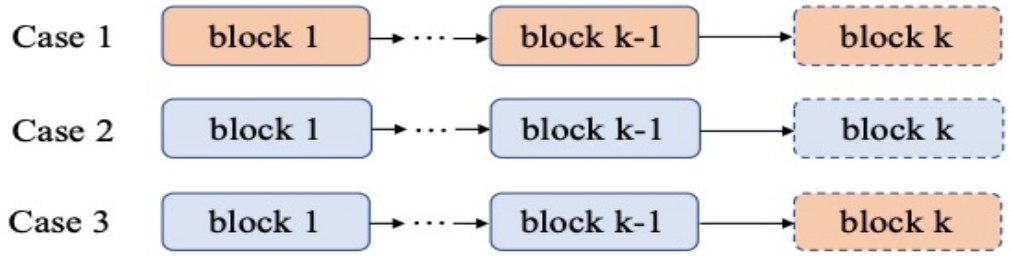
GDFQ [Xu et. al. SCUT. 2020]
ZAQ [Liu et. al. ECNU. 2020]


数据驱动的训练后量化
面向Transformer模型的训练后量化方法设计
QDrop [ICLR, 2022]
OS [NIPS, 2022]
OS+ [EMNLP, 2023]
大模型问世后，对量化方法提出了新的要求挑战
大模型难以实现无损量化，需求更高效微调方法
IR-QLoRA [ICML, 2024]
JSQ [ICML, 2024]
联合使用稀疏化和量化方法,进一步提高压缩率
在其他领域中应用量化技术
TMFQ-DM [CVPR, 2024]
REG-PTQ [CVPR, 2024]
PTQ4SAM [CVPR, 2024]
大模型压缩综述
PTSBench
LLMQBench
LLaMA3 Bench
为现有压缩方法提供评测基准，提供更多有价值的结论

Transformer量化顺序

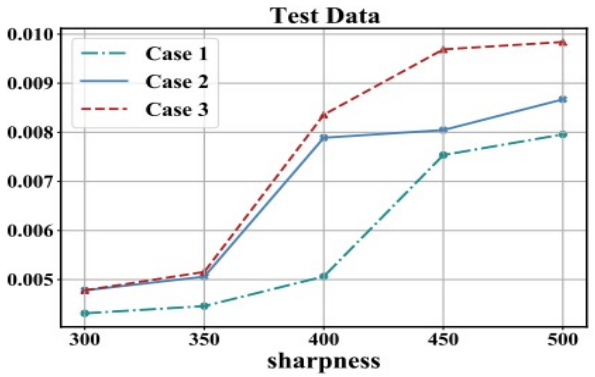
- **问题：现有微调权重的工作孤立了激活量化与权重微调顺序，引入激活值优化与过拟合问题存在冲突，无法均衡**

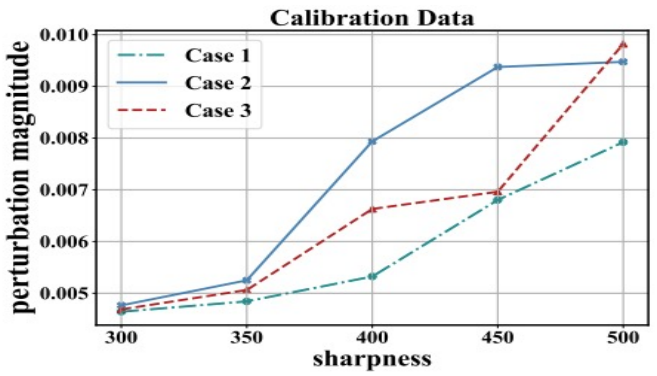
策略空间



精度对比

Case	1	2	3
ResNet-18	18.88	45.74	48.07
ResNet-50	4.34	46.98	49.07
MobileNetV2	5.83	50.71	51.20
RegNet-600MF	42.77	60.94	62.07
MnasNet	26.62	58.79	60.19





已有SOTA工作忽视激活值量化

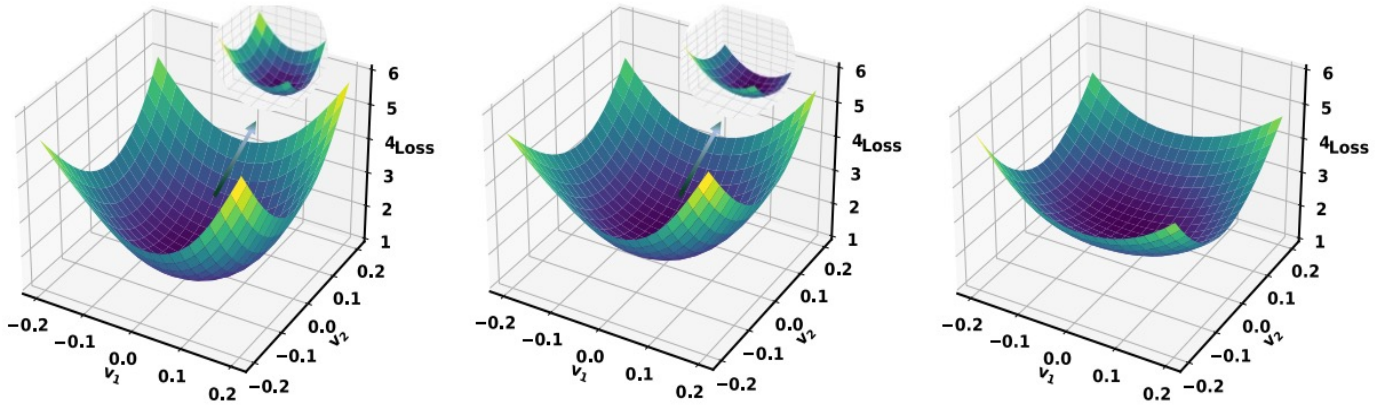
不同优化顺序对结果产生较大影响

$$\text{Case 1: } \boldsymbol{u}=\mathbf{0};\ \text{Case2: } \boldsymbol{u}=\frac{\hat{\boldsymbol{a}}}{\boldsymbol{a}}-1;\ \text{Case 3: } \boldsymbol{u}=\begin{cases}\frac{\hat{\boldsymbol{a}}}{\boldsymbol{a}}-1, & \text{block}_1 \sim \text{block}_{k-1}\\ \mathbf{0}, & \text{block}_k\end{cases}.$$

Wei et al., QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization, ICLR 2022

Transformer量化顺序

- **思路：引入合理激活值量化顺序和随机失活提高通用视角的平坦性，解决了离线量化过拟合问题，随机失活思想提升通用平坦性**



已有SOTA工作忽视激活值量化

↓

不合适的引入顺序会导致过拟合

↓

提高通用视角平坦性

↓

理论框架：在优化中量化激活值可以带来权重的平坦性

随机失活优化：概率性引入激活值量化

$$\boldsymbol{W}(\boldsymbol{a} \odot \begin{bmatrix} 1+\boldsymbol{u}_1(\boldsymbol{x}) \\ 1+\boldsymbol{u}_2(\boldsymbol{x}) \\ \dots \\ 1+\boldsymbol{u}_n(\boldsymbol{x}) \end{bmatrix}) = (\boldsymbol{W} \odot \begin{bmatrix} 1+\boldsymbol{u}_1(\boldsymbol{x}) & 1+\boldsymbol{u}_2(\boldsymbol{x}) & \dots & 1+\boldsymbol{u}_n(\boldsymbol{x}) \\ 1+\boldsymbol{u}_1(\boldsymbol{x}) & 1+\boldsymbol{u}_2(\boldsymbol{x}) & \dots & 1+\boldsymbol{u}_n(\boldsymbol{x}) \\ \dots & & & \\ 1+\boldsymbol{u}_1(\boldsymbol{x}) & 1+\boldsymbol{u}_2(\boldsymbol{x}) & \dots & 1+\boldsymbol{u}_n(\boldsymbol{x}) \end{bmatrix})\boldsymbol{a}.$$

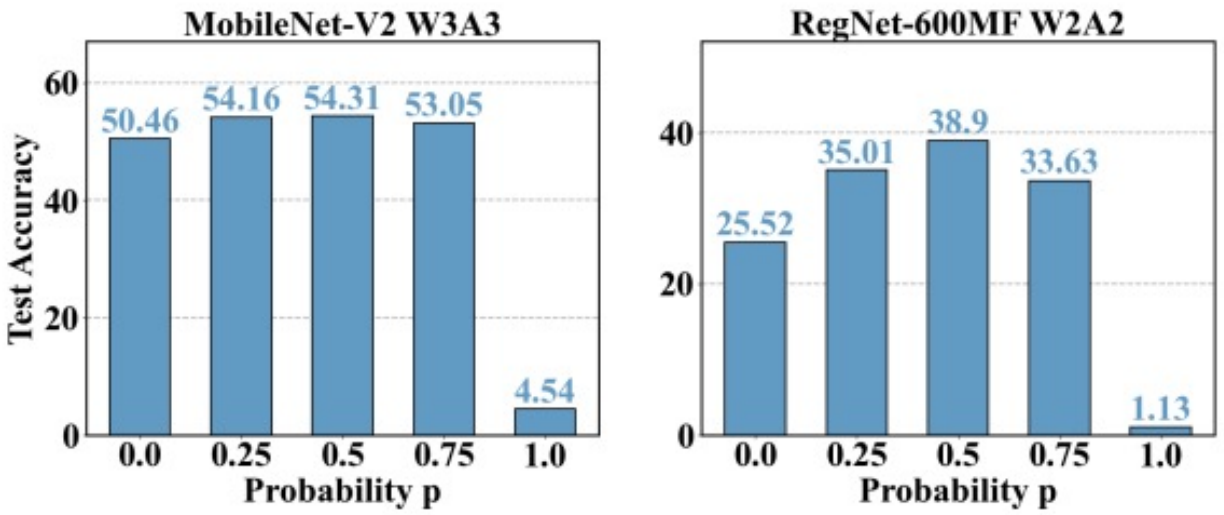
$$\text{QDROP}: u = \begin{cases} 0 & \text{with probability } p \\ \frac{\hat{a}}{a} - 1 & \text{with probability } 1-p \end{cases}.$$

Wei et al., QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization, ICLR 2022

Transformer量化顺序

■ 首次实现2比特离线量化精度可用，被集成至PaddleSlim等多个框架

方法	比特	Res18	Res50	MNV2	Reg600M	Reg3.2G
全精度	32/32	71.06	77.00	72.49	73.71	78.36
损失感受量化[84]	4/4	60.30	70.00	49.70	57.71*	55.89*
适应性量化法[85]	4/4	**69.60**	**75.90**	47.16*	-	-
比特划分法[86]	4/4	67.56	73.71	-	-	-
上下学习取整法[11]*	4/4	67.96	73.88	61.52	68.20	73.85
随机失活激活值量化	4/4	69.10	75.03	**67.89**	**70.62**	**76.33**
上下学习取整法 †*	4/4	69.36	74.76	64.33	-	-
块重构方法 †[14]	4/4	69.60	75.05	66.57	68.33	74.21
随机失活激活值量化 †	4/4	**69.62**	75.45	**68.84**	**71.18**	**76.66**
损失感受量化 *	2/4	0.18	0.14	0.13	0.17	0.12
适应性量化法 *	2/4	0.11	0.12	0.15	-	-
上下学习取整法 *	2/4	62.12	66.11	36.31	57.00	63.89
随机失活激活值量化	2/4	**64.66**	**70.08**	**52.92**	**63.10**	**70.95**
上下学习取整法 †*	2/4	64.14	68.40	41.52	59.27	65.33
块重构法 †	2/4	64.80	70.29	53.34	59.31	67.15
随机失活激活值量化 †	2/4	**65.25**	**70.65**	**54.22**	**63.80**	**71.70**

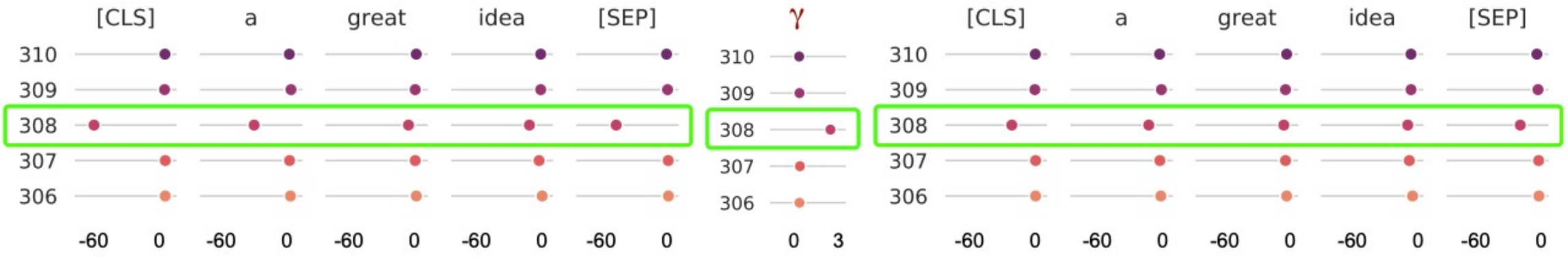


Method	Bits	ResNet-18	ResNet-50	MobileNet-V2	RegNet-600M	RegNet-3200M
BRECQ	w2a2	42.624	29.006	37.75 (w2a3)	3.332	3.728
Ours	w2a2	**54.710**	**58.260**	**42.01** (w2a3)	**41.068**	**54.410**
BRECQ	w4a4	**69.60**	75.05	66.57	68.33	74.21
Ours	w4a4	69.438	**75.476**	**68.464**	**71.252**	**76.718**

Wei et al., QDrop: Randomly Dropping Quantization for Extremely Low-bit Post-Training Quantization, ICLR 2022

LLM异常值抑制框架

■ 问题：大语言模型异常值呈现结构化特性，伽马放缩成为关键阻碍



$$\textbf{Non-scaling LayerNorm}: \quad X'_{t,j} = \frac{X_{t,j} - u_t}{\sqrt{\sigma_t^2 + \epsilon}} + \frac{\beta_j}{\gamma_j} \quad \Rightarrow \quad \textbf{LayerNorm}: \quad \widetilde{X}_{t,j} = \frac{X_{t,j} - u_t}{\sqrt{\sigma_t^2 + \epsilon}} \cdot \gamma_j + \beta_j$$

Tensor	0	1	2	3	4	5	6	7	8	9	10	11
$\widetilde{X}$	97.16	97.03	97.61	94.37	93.41	93.53	93.31	93.61	94.56	95.62	96.13	98.57
X'	**99.23**	**99.22**	**99.11**	**99.02**	**98.99**	**99.00**	**98.99**	**98.83**	**98.70**	**99.05**	**99.44**	**99.07**

LayerNorm伽马放大了异常值

部分异常值对精度贡献不大

Wei et al., Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models, NeurIPS 2022

LLM异常值抑制框架

- **思路：在大模型量化中引入的基于等价变换的异常值抑制框架，集成到多个框架**

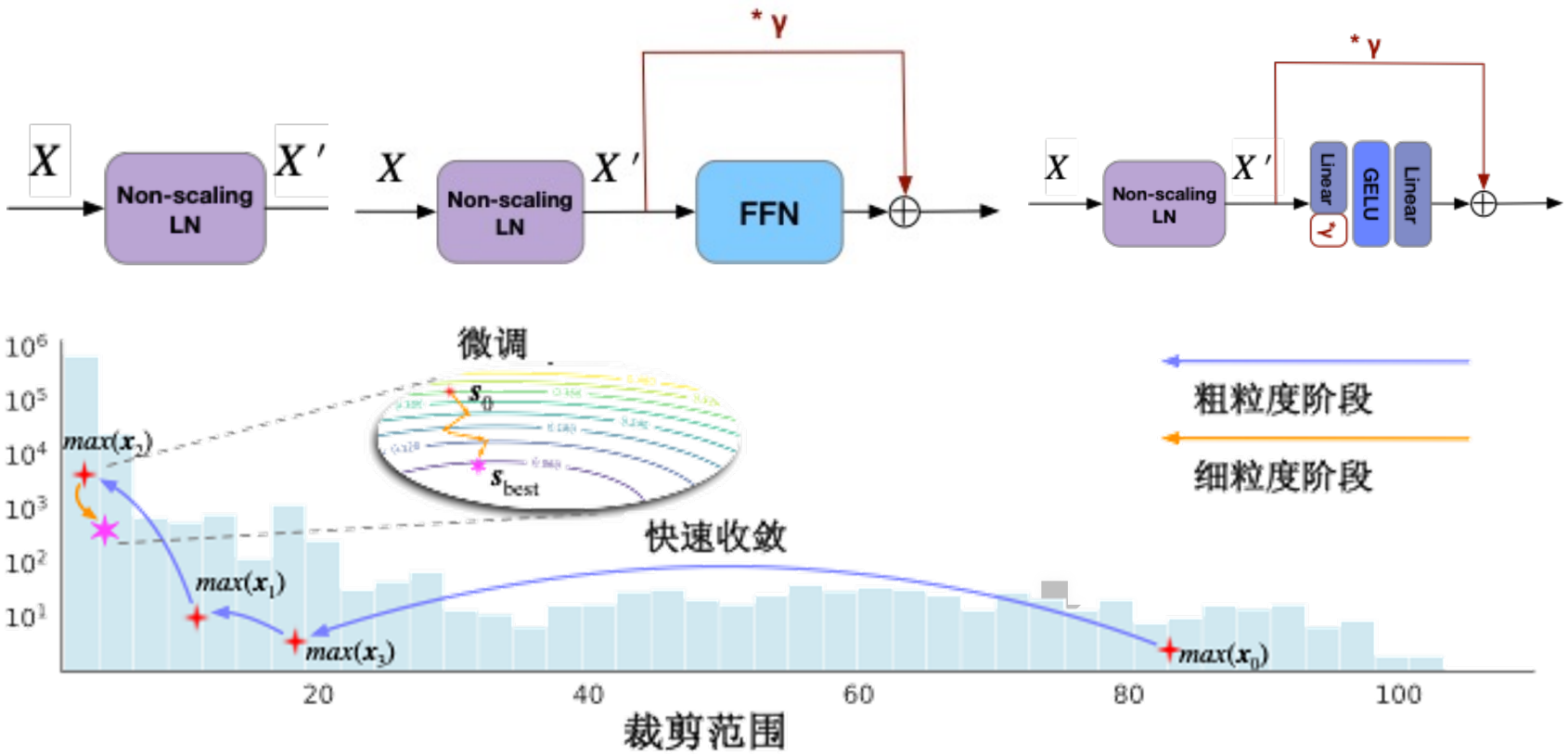
通道角度（伽马迁移）

$$X'_{t,j} = \frac{X_{t,j} - u_t}{\sqrt{\sigma_t^2 + \epsilon}} + \frac{\beta_j}{\gamma_j}$$

$$W(x \odot \begin{bmatrix}\gamma_1\\ \gamma_2\\ \cdots\\ \gamma_n\end{bmatrix}) = (W \odot \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_n\\ \gamma_1 & \gamma_2 & \cdots & \gamma_n\\ \cdots & & & \\ \gamma_1 & \gamma_2 & \cdots & \gamma_n\end{bmatrix})x$$

令牌角度（基于令牌的裁剪）

$$L(s) = \|\hat{f}(s) - f\|_F^2$$



Wei et al., Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models, NeurIPS 2022

LLM异常值抑制框架

■ 首次将语言模型的离线量化做到6比特精度可用

摘要任务

Method	Bits	BERT		RoBERTa		BART	
	(W-E-A)	SQuAD v1.1	SQuAD v2.0	SQuAD v1.1	SQuAD v2.0	SQuAD v1.1	SQuAD v2.0
Full Prec.	32-32-32	88.28/80.82	77.34/73.60	92.25/85.83	83.30/80.26	91.63/84.79	80.82/77.41
OMSE [28]	8-8-8	**87.90/80.16**	76.88/73.08	91.48/84.53	82.53/79.41	90.49/83.11	79.62/76.12
Ours	8-8-8	87.60/79.80	**76.93/73.14**	**91.57/84.86**	**82.94/79.72**	**91.08/84.07**	**80.55/77.04**
OMSE	6-6-6	79.77/69.10	67.52/63.09	70.64/58.80	45.80/39.95	81.44/70.61	67.89/63.29
Percentile [29]	6-6-6	78.55/67.14	69.12/65.64	67.24/53.28	56.38/51.58	82.45/72.87	68.44/63.29
EasyQuant [40]	6-6-6	80.47/70.08	71.95/68.06	67.85/55.92	47.99/42.21	82.41/71.72	69.93/64.94
Ours	6-6-6	**84.48/75.53**	**74.69/70.55**	**80.79/70.83**	**68.47/64.10**	**83.68/75.34**	**74.44/70.36**

问答任务

Method	Bits	BERT		RoBERTa		BART	
	(W-E-A)	SQuAD v1.1	SQuAD v2.0	SQuAD v1.1	SQuAD v2.0	SQuAD v1.1	SQuAD v2.0
Full Prec.	32-32-32	88.28/80.82	77.34/73.60	92.25/85.83	83.30/80.26	91.63/84.79	80.82/77.41
OMSE [28]	8-8-8	**87.90/80.16**	76.88/73.08	91.48/84.53	82.53/79.41	90.49/83.11	79.62/76.12
Ours	8-8-8	87.60/79.80	**76.93/73.14**	**91.57/84.86**	**82.94/79.72**	**91.08/84.07**	**80.55/77.04**
OMSE	6-6-6	79.77/69.10	67.52/63.09	70.64/58.80	45.80/39.95	81.44/70.61	67.89/63.29
Percentile [29]	6-6-6	78.55/67.14	69.12/65.64	67.24/53.28	56.38/51.58	82.45/72.87	68.44/63.29
EasyQuant [40]	6-6-6	80.47/70.08	71.95/68.06	67.85/55.92	47.99/42.21	82.41/71.72	69.93/64.94
Ours	6-6-6	**84.48/75.53**	**74.69/70.55**	**80.79/70.83**	**68.47/64.10**	**83.68/75.34**	**74.44/70.36**

6比特精度对齐浮点

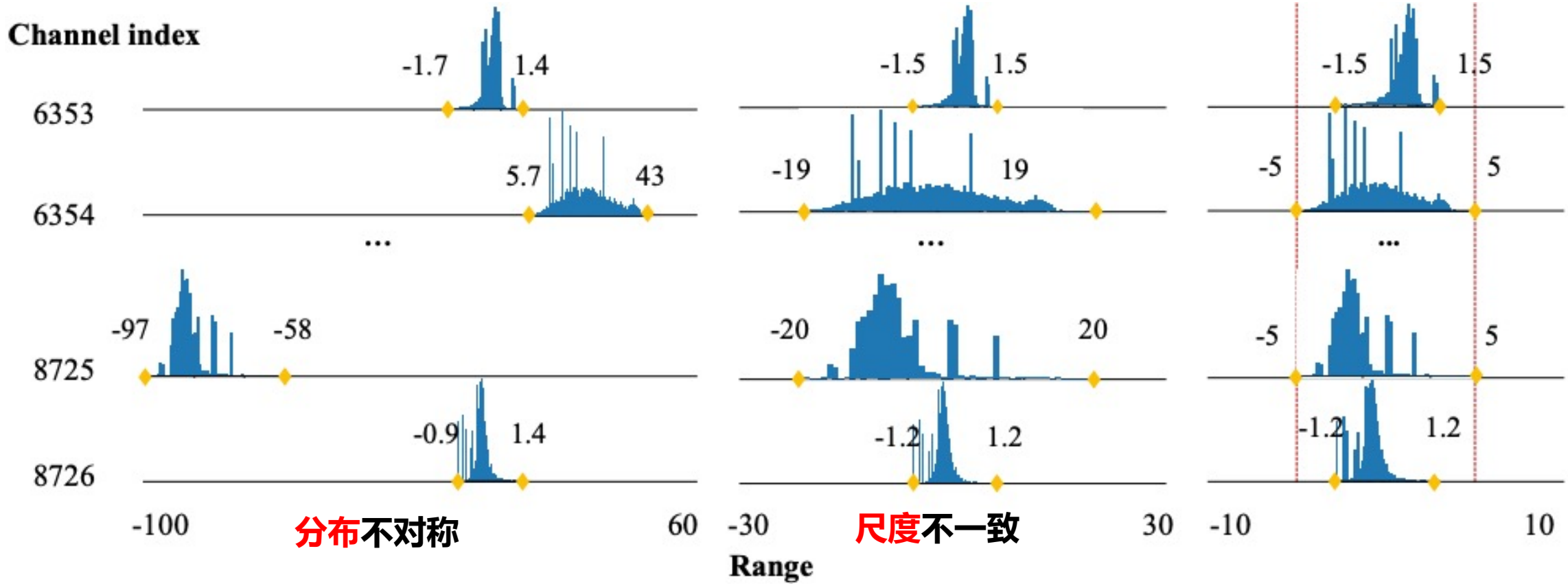
在极低比特下，大幅度超过当前SOTA

对量化敏感模型效果明显

Wei et al., Outlier Suppression: Pushing the Limit of Low-bit Transformer Language Models, NeurIPS 2022

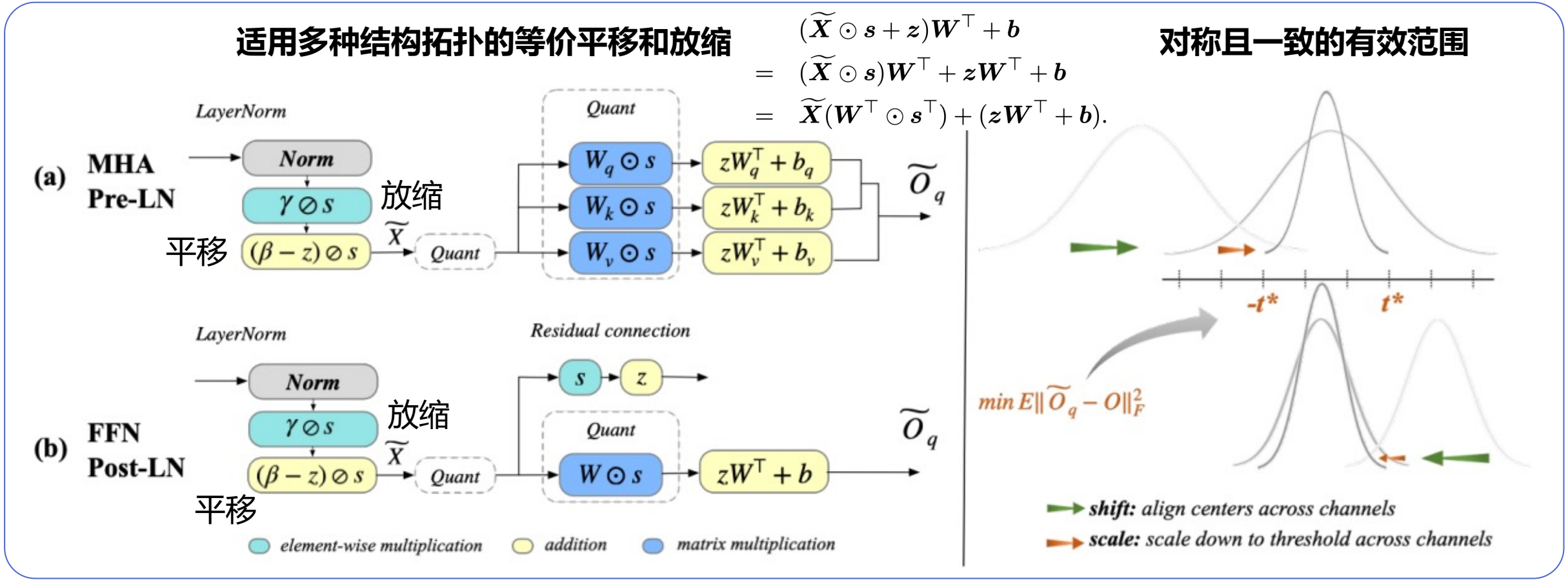
LLM异常值抑制框架+

■ 问题：大语言模型异常值存在通道不对称性和尺度不一致性



LLM异常值抑制框架+

■ 思路：通过通道维度的平移缩放抑制结构化异常值和非对称分布

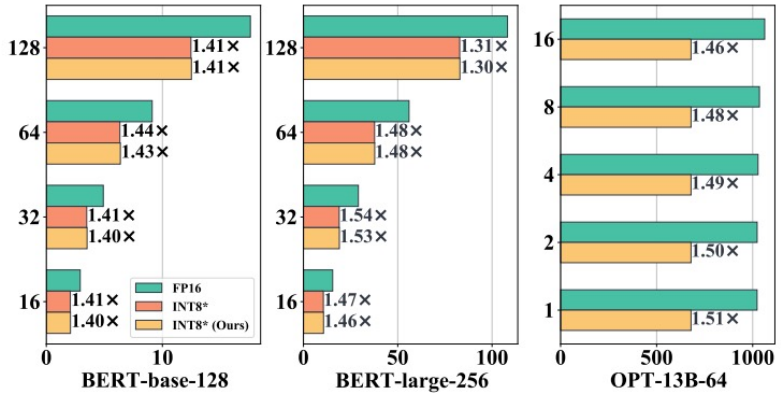


LLM异常值抑制框架+

■ 有效抑制异常值范围，首次实现INT4可用精度

INT4可用SOTA效果

Model	Method	PIQA (↑)			Winogrande (↑)			HellaSwag (↑)			WikiText2 (↓)		
		FP16	INT6	INT4	FP16	INT6	INT4	FP16	INT6	INT4	FP16	INT6	INT4
LLaMA-1-7B	MinMax	77.37	77.26	55.98	66.93	66.54	49.64	72.99	71.78	32.28	5.68	6.00	473.97
	SmoothQuant		77.18	70.08		65.51	52.96		72.10	58.13		5.85	16.87
	OS+		**77.48**	**72.31**		**67.01**	**56.67**		**72.32**	**61.24**		**5.76**	**14.17**
LLaMA-1-13B	MinMax	79.05	78.56	50.65	70.09	69.53	50.28	76.22	75.26	26.34	5.09	5.58	3410.45
	SmoothQuant		78.45	66.49		**69.69**	51.78		75.20	58.95		5.25	56.75
	OS+		**78.73**	**75.03**		69.53	**61.17**		**75.74**	**67.21**		**5.22**	**18.95**
LLaMA-1-30B	MinMax	80.09	78.40	50.00	72.77	72.45	50.12	79.21	77.25	27.09	4.10	5.09	2959.15
	SmoothQuant		78.78	71.55		73.01	54.54		78.13	60.97		4.40	51.47
	OS+		**79.98**	**73.01**		**73.64**	**60.38**		**78.77**	**68.03**		**4.30**	**22.61**
LLaMA-1-65B	MinMax	80.85	77.58	50.27	77.11	69.46	49.33	80.73	78.72	24.59	3.56	5.25	14584.66
	SmoothQuant		78.40	65.02		74.30	51.14		78.57	59.78		3.77	19.37
	OS+		**80.47**	**74.43**		**75.14**	**61.72**		**79.76**	**67.65**		**3.65**	**9.33**



Method	activation		weight		Output change
	range	MSE	range	MSE	MSE
original	(-93.9, 31.6)	209.8	(-0.13, 0.13)	0.001	18061.5
OS	(-23.5,15.7)	142.9	(-0.40, 0.41)	**0.006**	6182.52
SQ	(-3.5, 2.0)	**3.65**	(3.4, 3.5)	0.43	3535.86
Our scaling	(-8.4, 8.4)	48.54	(1.2, 1.3)	0.02	**1334.89**

Wei et al., Outlier Suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling, EMNLP 2023

LLM高效微调量化

■ 问题：现有的LoRA微调量化方法会导致量化后的LLM严重退化

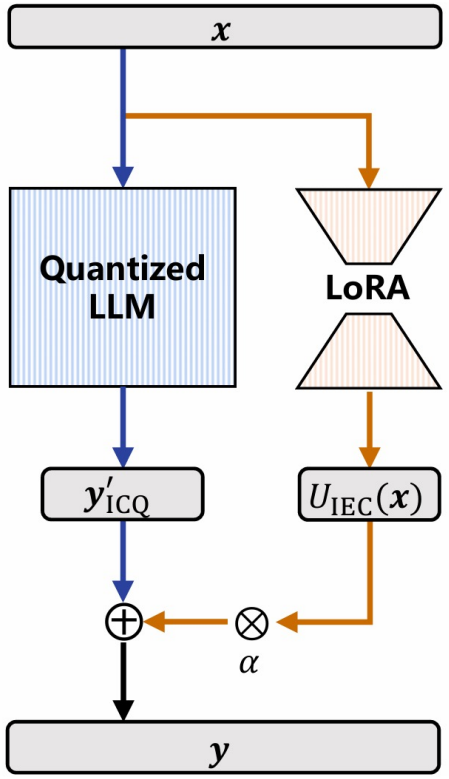
Q1：LLM的NF量化是否准确？

$$\mathcal{I}(\hat{\boldsymbol{w}}^{\text{FP16}}; \boldsymbol{w}) = \mathcal{H}(\hat{\boldsymbol{w}}^{\text{FP16}}) - \underbrace{\mathcal{H}(\hat{\boldsymbol{w}}^{\text{FP16}} \mid \boldsymbol{w})}_{0},$$

$$\underset{s, s_1^{\text{FP8}}, s_2^{\text{FP16}}}{argmax}\, \mathcal{H}(\hat{\boldsymbol{w}}^{\text{FP16}}; s, s_1^{\text{FP8}}, s_2^{\text{FP16}}). \quad (6)$$

$$\underset{s}{argmax}\, \mathcal{H}(\hat{\boldsymbol{w}}^{\text{NF}k}; s) = -\sum_{i=1}^{2^k-1} P(q_i) \log_2 P(q_i), \quad (7)$$

量化过程的位宽减小
限制了表达能力



Q2：LoRA的低秩是否存在局限性？

$$\boldsymbol{y} = \boldsymbol{y}' + \alpha \boldsymbol{x} \boldsymbol{\ell}_1 \boldsymbol{\ell}_2$$

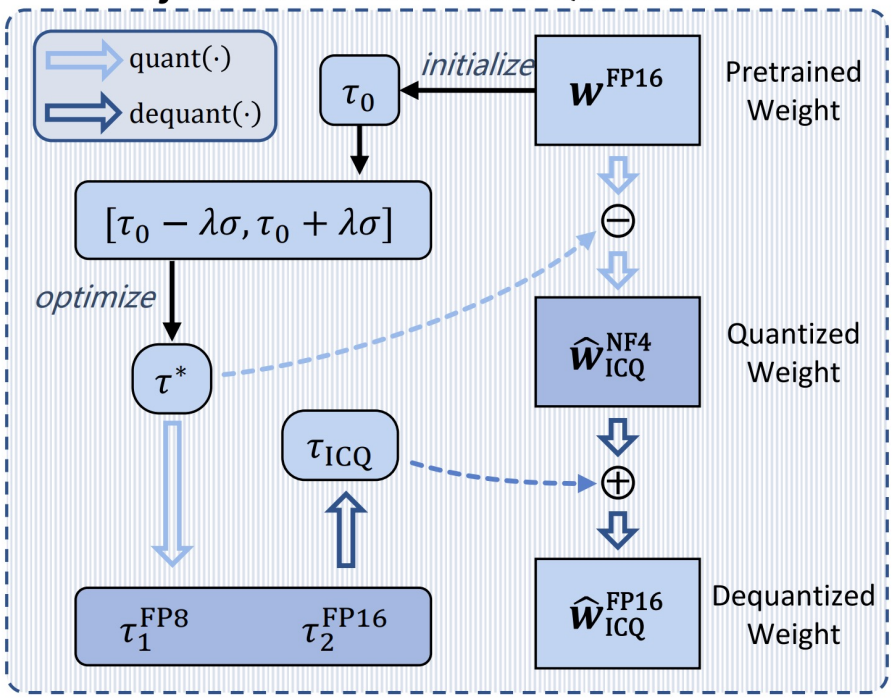
$$\boldsymbol{\ell}_1 \in \mathbb{R}^{h \times r} \text{ and } \boldsymbol{\ell}_2 \in \mathbb{R}^{r \times o}$$

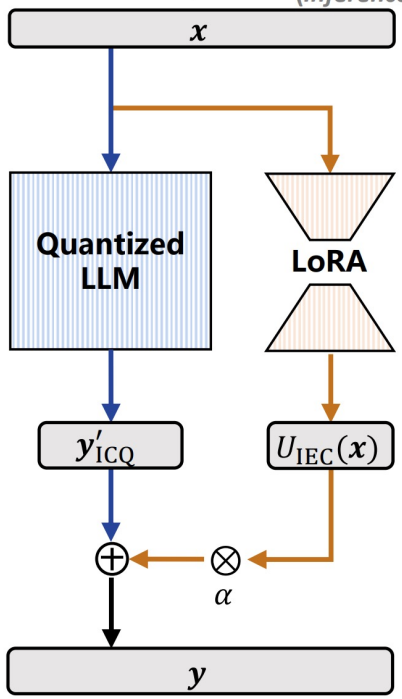
LoRA微调过程中的秩非常低
(r ≪ *h*, *o*)

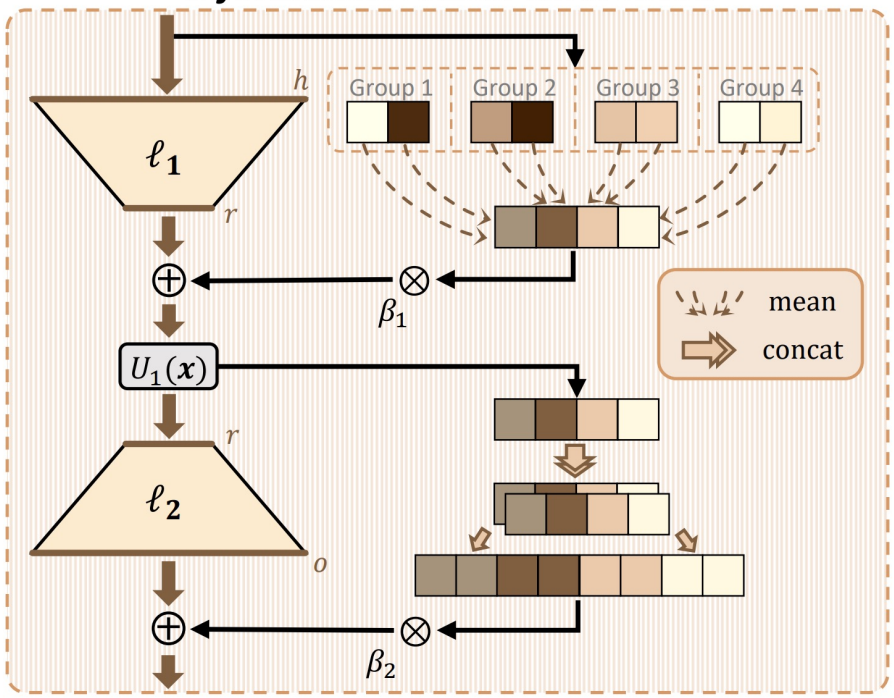
低秩限制了表达能力

LLM高效微调量化

■ 思路：从信息角度，解决大模型LoRA微调量化方法的信息损失问题







$$\hat{\boldsymbol{w}}_{\text{ICQ}}^{\text{NF}k} = \text{NF}k\left(\frac{\boldsymbol{w}-\tau^*}{\text{absmax}(\boldsymbol{w}-\tau^*)}\right),$$

$$\hat{\boldsymbol{w}}_{\text{ICQ}}^{\text{FP16}} = \hat{\boldsymbol{w}}_{\text{ICQ}}^{\text{NF}k}\,\text{dequant}(s_1^{\text{FP8}}, s_2^{\text{FP16}}) + \text{dequant}(\tau_1^{\text{FP8}}, \tau_2^{\text{FP16}}).$$

$$U_1(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\ell}_1 + \beta_1 \prod^{\frac{r}{\gcd(h,r)}} \left(\frac{\gcd(h,r)}{h}\sum_{i=1}^{\gcd(h,r)} \boldsymbol{x}^{\left[(i-1)\frac{h}{\gcd(h,r)}:i\frac{h}{\gcd(h,r)}-1\right]}\right),$$

$$U_2(\boldsymbol{x}') = \boldsymbol{x}'\boldsymbol{\ell}_2 + \beta_2 \prod^{\frac{o}{\gcd(o,r)}} \left(\frac{\gcd(o,r)}{r}\sum_{i=1}^{\gcd(o,r)} \boldsymbol{x}'^{\left[(i-1)\frac{r}{\gcd(o,r)}:i\frac{r}{\gcd(o,r)}-1\right]}\right),$$

信息校准量化使得LLM的量化参数能够准确地保留原始信息

信息保留
准确高效微调

信息弹性连接使LoRA能够使用具有多样化信息的弹性表示

Qin et al., Accurate LoRA-Finetuning Quantization of LLMs via Information Retention, ICML 2024(Oral)

LLM高效微调量化

IR-QLoRA以很小的代价实现了高效微调结果的显著提升

Method	#Bit	MMLU				
		Hums.	STEM	Social	Other	Avg.
LLaMA-7B	16	33.3	29.8	37.8	38.0	34.6
PEQA	4	34.9	28.9	37.5	40.1	34.8
NormalFloat	4	33.1	30.6	38.8	38.8	35.1
QLoRA w/ GPTQ	4	33.8	31.3	37.4	42.2	36.0
QLoRA	4	36.1	31.9	42.0	44.5	38.4
QA-LoRA	4	36.6	32.4	44.8	44.9	39.4
IR-QLoRA (ours)	4	38.6	34.6	45.2	45.5	**40.8**
LLaMA-13B	16	40.6	36.7	48.9	48.0	43.3
NormalFloat	4	43.0	34.5	51.8	51.4	45.0
PEQA	4	43.0	37.7	53.6	49.0	45.0
QLoRA	4	45.4	37.4	55.7	54.3	48.0
QLoRA w/ GPTQ	4	48.4	38.3	54.9	55.2	49.2
QA-LoRA	4	48.4	38.3	54.9	55.2	49.2
IR-QLoRA (ours)	4	47.2	39.0	56.5	55.0	**49.3**
LLaMA-30B	16	56.2	45.9	67.1	63.9	58.2
NormalFloat	4	55.3	44.7	66.2	63.3	57.3
QLoRA	4	55.4	46.0	66.4	63.6	57.7
QLoRA w/ GPTQ	4	55.8	46.4	67.0	64.0	58.1
QA-LoRA	4	55.8	46.4	67.0	64.0	58.1
IR-QLoRA (ours)	4	56.7	46.7	66.5	63.2	**58.2**
LLaMA-65B	16	61.4	51.9	73.6	67.6	63.4
QA-LoRA	4	60.8	50.5	72.5	66.7	62.5
NormalFloat	4	60.7	52.3	72.6	67.3	63.0
QLoRA w/ GPTQ	4	60.4	52.5	73.0	67.2	63.0
QLoRA	4	60.3	52.7	72.9	67.4	63.1
IR-QLoRA (ours)	4	60.1	50.1	74.4	68.7	**63.1**

Alpaca数据集微调后的4 bit性能对比

Method	#Bit	MMLU				
		Hums.	STEM	Social	Other	Avg.
LLaMA-7B	16	33.3	29.8	37.8	38.0	34.6
NormalFloat	4	33.1	30.6	38.8	38.8	35.1
QLoRA w/ GPTQ	4	33.8	31.3	37.4	42.2	36.0
QLoRA	4	41.4	35.0	49.8	52.0	44.3
QA-LoRA	4	43.9	38.0	54.3	53.0	47.0
IR-QLoRA (ours)	4	44.2	39.3	54.5	52.9	**47.4**
LLaMA-13B	16	40.6	36.7	48.9	48.0	43.3
NormalFloat	4	43.0	34.5	51.8	51.4	45.0
QLoRA w/ GPTQ	4	48.4	38.3	54.9	55.2	49.2
QLoRA	4	49.9	40.1	60.2	57.9	51.9
QA-LoRA	4	50.0	41.5	60.5	58.4	52.4
IR-QLoRA (ours)	4	49.2	41.2	62.1	59.2	**52.6**
LLaMA-30B	16	56.2	45.9	67.1	63.9	58.2
NormalFloat	4	55.3	44.7	66.2	63.3	57.3
QLoRA w/ GPTQ	4	55.8	46.4	67.0	64.0	58.1
QLoRA	4	57.2	48.6	69.8	65.2	60.0
QA-LoRA	4	57.9	48.8	71.0	65.5	60.6
IR-QLoRA (ours)	4	58.1	49.4	70.7	65.8	**60.8**
LLaMA-65B	16	61.4	51.9	73.6	67.6	63.4
NormalFloat	4	60.7	52.3	72.6	67.3	63.0
QLoRA w/ GPTQ	4	60.4	52.5	73.0	67.2	63.0
QLoRA	4	59.8	52.9	75.0	69.6	63.9
QA-LoRA	4	57.6	51.1	73.9	67.4	62.1
IR-QLoRA (ours)	4	61.6	52.0	75.6	68.9	**64.3**

Flan v2数据集微调后的4 bit性能对比

Method	Data	#Bit	MMLU				
			Hums.	STEM	Social	Other	Avg.
LLaMA-7B	-	16	33.3	29.8	37.8	38.0	34.6
NormalFloat	-	3	30.5	29.9	34.8	34.9	32.3
QLoRA w/ GPTQ	Alpaca	3	31.6	30.1	35.6	39.8	34.0
QLoRA	Alpaca	3	35.8	32.1	40.7	43.1	37.8
QA-LoRA	Alpaca	3	35.6	30.5	41.5	42.7	37.4
IR-QLoRA	Alpaca	3	36.0	33.9	42.2	42.7	**38.4**
QLoRA w/ GPTQ	Flan v2	3	32.2	31.7	42.7	42.8	36.9
QLoRA	Flan v2	3	41.3	37.1	50.9	49.8	44.5
QA-LoRA	Flan v2	3	41.3	36.0	52.8	50.2	44.7
IR-QLoRA	Flan v2	3	43.0	37.7	52.3	51.7	**45.9**
NormalFloat	-	2	24.2	28.9	31.1	25.0	26.9
QLoRA w/ GPTQ	Alpaca	2	23.4	26.2	26.4	28.4	25.8
QLoRA	Alpaca	2	24.0	27.0	27.5	26.7	26.2
QA-LoRA	Alpaca	2	27.3	26.1	26.1	30.3	27.5
IR-QLoRA	Alpaca	2	26.0	27.8	30.2	28.3	**27.8**
QLoRA w/ GPTQ	Flan v2	2	23.9	25.3	26.2	25.3	25.0
QLoRA	Flan v2	2	31.8	28.7	36.7	37.7	33.5
QA-LoRA	Flan v2	2	31.8	28.1	34.5	38.5	33.2
IR-QLoRA	Flan v2	2	31.7	29.4	37.8	36.5	**33.7**

2-3 bit下的性能对比

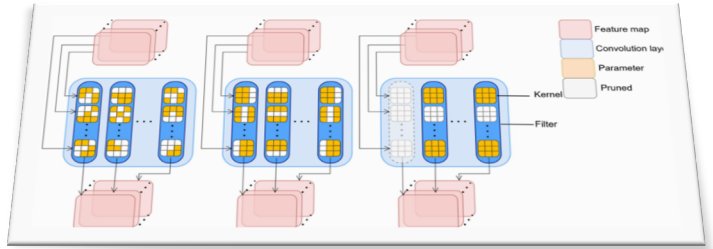
在2~4 bit位宽下显著提高 LLaMA 和 LLaMA2 系列高效微调的准确率
4-bit LLaMA-7B在MMLU上比现有方法提高了1.4%，并仅需0.31%的额外时间

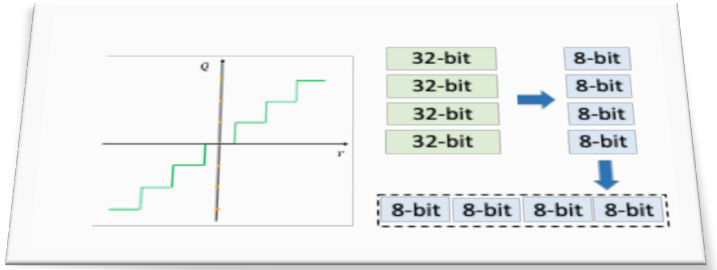
LLM稀疏量化联合压缩

- **问题：单独使用量化方法难以在极端压缩率下保持模型性能，需要考虑稀疏量化联合场景下，对异常值的耦合处理问题**

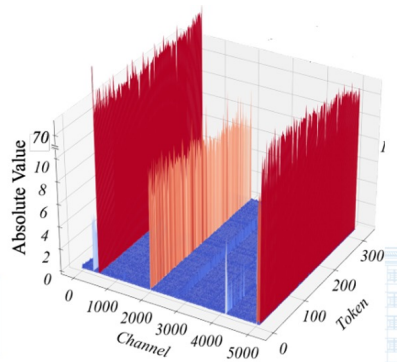
挑战1 稀疏化量化在大模型场景中耦合性差

稀疏倾向于保留绝对值更大的参数



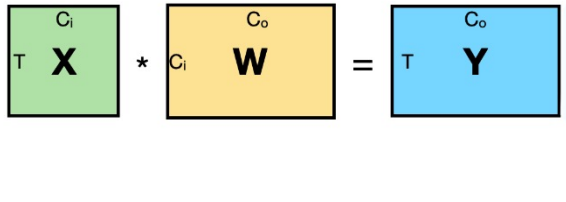


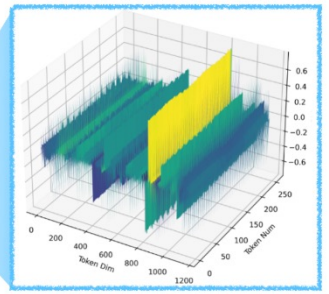
量化偏好变化范围更小的参数

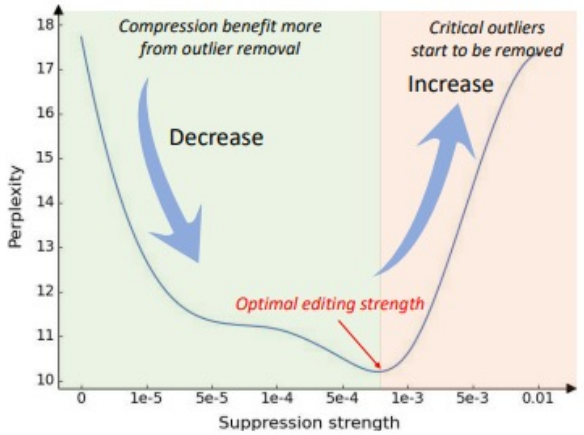


?

挑战2 大模型中存在大量异常值







Guo et al., Compressing Large Language Models by Joint Sparsification and Quantization, ICML 2024

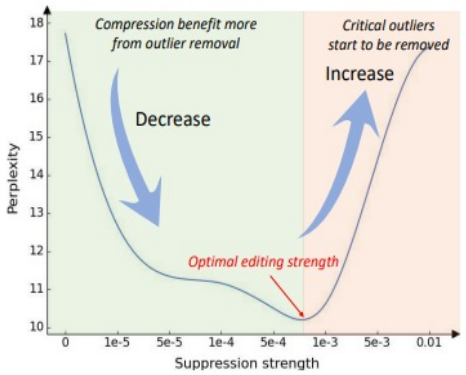
LLM稀疏量化联合压缩

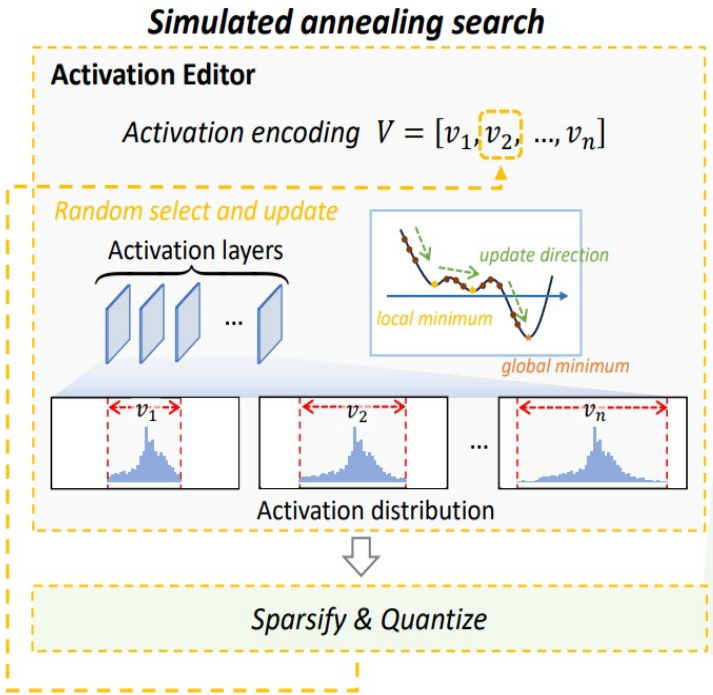
■ 思路：提出联合使用稀疏化与量化，综合利用两种方法的优势

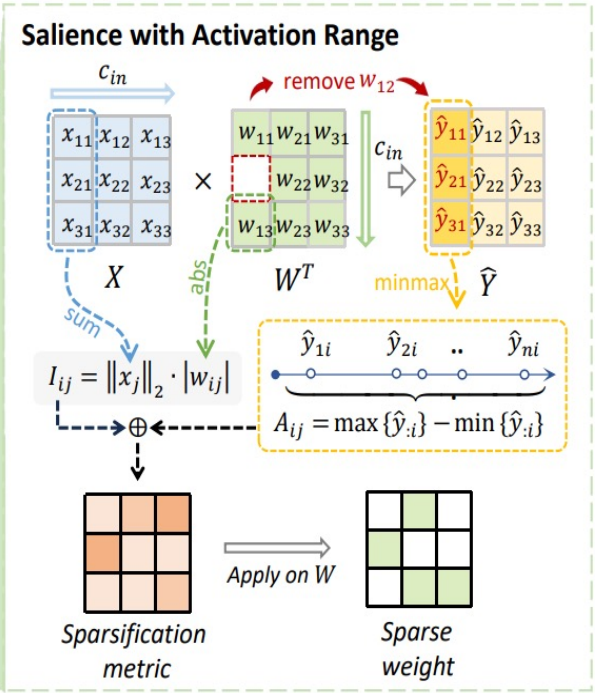
基于搜索的激活编辑器启发式的对异常值进行处理 ⟵ 稀疏量化联合框架 ⟶ **激活范围显著性度量**能够平衡量化和剪枝权衡

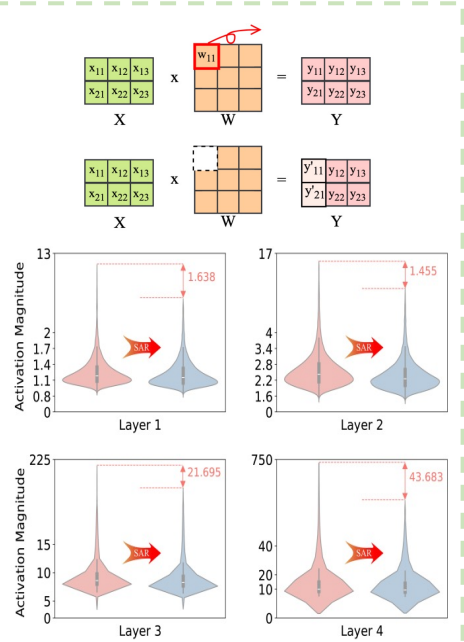
$$\mathbf{F}_{edit} = clamp(\mathbf{F}, LB, HB),$$
$$where \quad LB = r \cdot max(\mathbf{F}) + (1-r) \cdot min(\mathbf{F}),$$
$$and \quad HB = (1-r) \cdot max(\mathbf{F}) + r \cdot min(\mathbf{F}).$$

$$V^* = \arg\max_V \mathscr{P}(COM(LLM; V))$$









Guo et al., Compressing Large Language Models by Joint Sparsification and Quantization, ICML 2024

LLM稀疏量化联合压缩

- **在所有数据集和模型取得最好结果，带来33%推理加速**

零样本生成表现评估

方法	计算量（#MACs）	模型大小	WikiText↓	PIQA	BoolQ	MMLU	HellaSwag	Arc-e	Arc-c	WinoGrande	平均值
LLaMA-7B[6]	14.64T	12.6G	9.42	79.16	74.95	33.0	76.20	63.59	44.71	70.01	63.09
Wanda[12]	3.24T	1.8G	260364.00	50.98	48.41	22.9	26.46	26.85	26.88	50.28	36.11
LLM-Pruner[13]	3.24T	1.8G	2535837.61	49.35	46.91	23.5	26.26	26.77	28.92	49.96	35.95
SparseGPT[33]	3.24T	1.8G	5789.33	51.41	37.98	23.2	27.15	26.01	25.09	48.70	34.22
SmoothQuant[19]	3.24T	4.7G	12.70	76.93	72.60	28.4	71.56	70.16	41.72	66.06	61.06
OmniQuant[1]	3.24T	4.7G	12.49	77.24	73.27	28.5	71.33	69.10	42.11	68.10	61.38
Wanda+SmoothQuant	3.24T	3.5G	12.34	77.97	72.69	28.1	71.84	68.27	41.89	68.43	61.31
SparseGPT+SmoothQuant	3.24T	3.5G	11.98	77.20	73.24	30.2	70.91	68.48	41.21	69.30	61.51
JSQ (Ours)	3.24T	3.5G	11.04	78.72	74.58	30.0	72.33	69.30	43.42	69.58	62.56
LLaMA-13B[6]	26.52T	24.3G	8.21	80.14	77.89	42.1	79.09	74.75	47.70	72.77	67.78
Wanda[12]	4.09T	3.4G	172946.83	50.33	37.83	25.7	26.47	26.81	26.54	49.49	34.74
LLM-Pruner[13]	4.09T	3.4G	934523.91	49.67	38.10	25.0	25.52	25.97	29.10	49.96	34.76
SparseGPT[33]	4.09T	3.4G	1265.67	51.36	38.07	23.2	28.02	27.82	23.46	49.72	34.52
SmoothQuant[19]	4.09T	9.1G	11.65	77.80	70.98	31.9	76.97	68.52	41.72	69.61	62.50
OmniQuant[1]	4.09T	9.1G	10.27	78.80	75.43	36.4	75.31	70.11	43.63	70.32	64.29
Wanda+SmoothQuant	4.09T	6.8G	10.39	78.45	75.66	31.5	75.54	71.21	44.71	69.61	63.81
SparseGPT+SmoothQuant	4.09T	6.8G	10.49	78.40	76.61	32.8	75.45	69.53	44.11	67.96	63.55
JSQ (Ours)	4.09T	6.8G	9.58	80.05	78.16	38.2	75.41	72.93	46.67	72.24	66.24
LLaMA-33B[6]	33.12T	60.6G	6.24	82.21	82.72	43.1	82.62	78.91	52.90	75.77	71.18
Wanda[12]	4.77T	8.5G	313764.56	50.00	42.78	22.8	26.20	26.56	27.39	49.64	35.05
LLM-Pruner[13]	4.77T	8.5G	3533383.20	48.97	39.63	24.2	25.94	26.18	26.96	49.01	34.41
SparseGPT[33]	4.77T	8.5G	461.82	52.34	57.25	22.9	29.10	28.70	22.18	47.99	37.21
SmoothQuant[19]	4.77T	22.7G	11.05	77.04	76.24	36.8	76.86	72.56	50.17	72.45	66.02
OmniQuant[1]	4.77T	22.7G	9.41	79.05	78.20	45.1	78.54	73.37	50.82	75.22	68.61
Wanda+SmoothQuant	4.77T	17.0G	8.00	77.97	80.98	38.2	79.35	74.71	51.11	74.19	68.07
SparseGPT+SmoothQuant	4.77T	17.0G	8.05	77.75	82.78	43.7	79.20	74.37	50.43	74.66	68.98
JSQ (Ours)	4.77T	17.0G	7.68	79.20	81.92	49.3	80.76	76.35	51.92	77.51	70.27
LLaMA-2-7B[7]	14.64T	12.6G	8.79	79.11	77.71	41.6	76.01	74.49	46.25	69.06	66.32
Wanda[12]	3.24T	1.8G	100584.54	48.97	37.83	25.5	26.31	26.26	26.96	49.96	34.54
SmoothQuant[19]	3.24T	4.7G	12.22	76.12	72.32	35.2	72.86	70.63	43.69	64.56	62.20
Wanda+SmoothQuant	3.24T	3.5G	10.74	78.35	75.44	34.2	72.91	69.53	43.17	67.17	62.97
JSQ (Ours)	3.24T	3.5G	10.64	79.09	76.10	34.6	73.28	71.23	45.56	69.03	64.13
LLaMA2-13B[7]	26.52T	24.3G	7.90	80.52	80.61	52.1	79.37	77.48	49.06	72.30	70.21
Wanda[12]	4.09T	3.4G	201702.58	49.29	37.83	22.9	25.81	27.06	26.79	51.70	34.48
SmoothQuant[19]	4.09T	9.1G	9.32	77.97	76.42	46.4	75.85	73.78	45.14	67.64	66.17
Wanda+SmoothQuant	4.09T	6.8G	9.79	78.54	78.91	46.7	76.64	73.23	47.61	70.09	67.38
JSQ (Ours)	4.09T	6.8G	8.58	79.25	79.61	48.5	76.32	75.51	46.25	71.84	68.18
ChatGLM3-6B[24]	12.24T	11.6G	10.12	81.45	86.45	62.1	78.01	79.59	53.58	72.61	73.40
Wanda[12]	2.20T	1.6G	49249.03	51.58	49.20	23.0	26.18	25.00	25.94	49.33	35.75
SmoothQuant[19]	2.20T	4.4G	15.23	73.12	75.54	48.7	51.30	60.82	36.60	53.83	57.13
Wanda+SmoothQuant	2.20T	3.3G	12.01	78.67	83.43	57.7	72.42	73.86	46.59	68.35	68.72
JSQ (Ours)	2.20T	3.3G	11.83	79.89	83.89	57.9	73.34	75.10	47.02	69.74	69.55

硬件加速性能

表 6 WikiText 上 2:4 和 4:8 半结构化稀疏性能

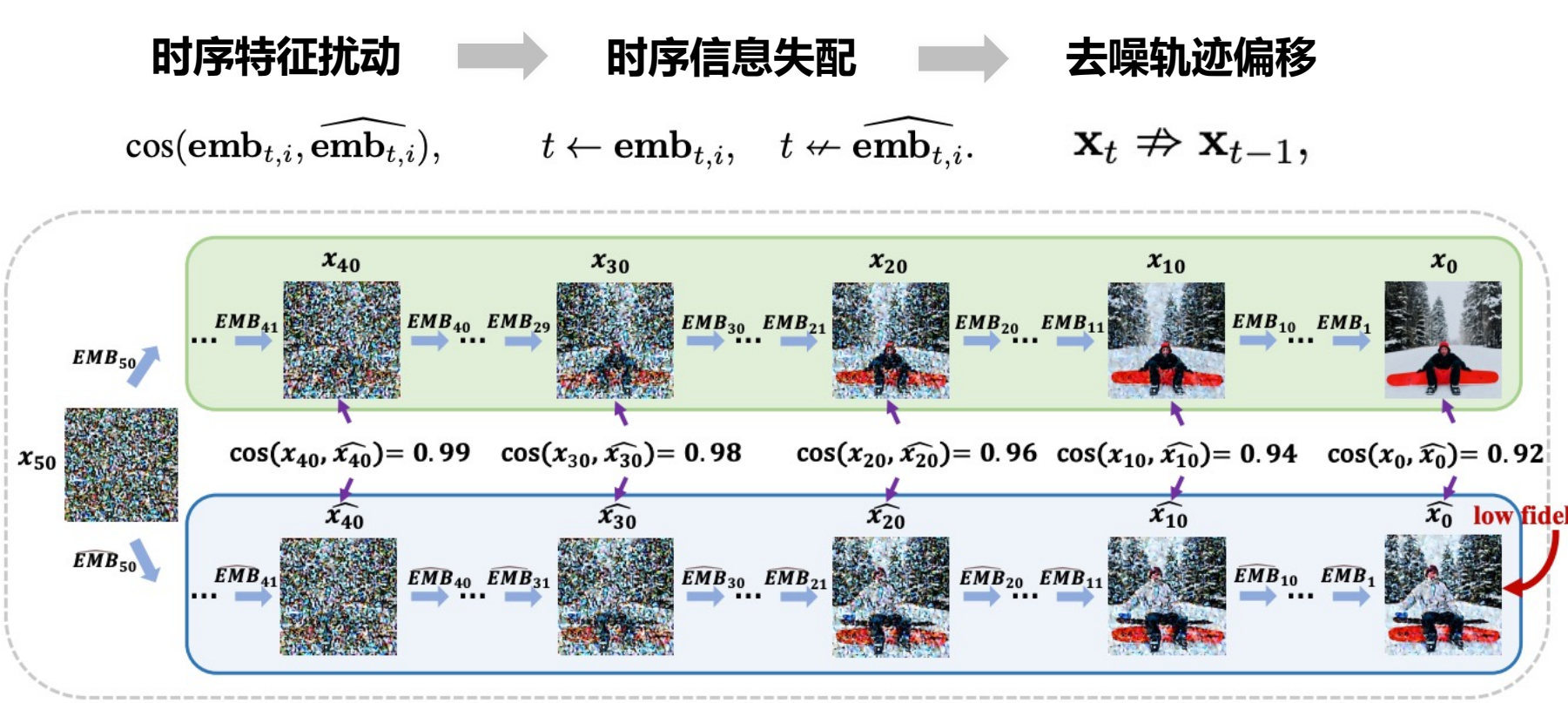
方法	计算量（#MACs）	模型大小	WikiText↓
LLaMA-7B	14.64T	12.6G	9.42
Unstructured	3.03T	3.2G	11.45
Structured 2:4	3.03T	3.2G	13.54
Structured 4:8	3.03T	3.2G	12.15
LLaMA2-7B	14.64T	12.6G	8.79
Unstructured	3.03T	3.2G	10.89
Structured 2:4	3.03T	3.2G	13.24
Structured 4:8	3.03T	3.2G	12.07

在相同计算量下，JSQ 框架在几乎所有数据集和模型上取得最好结果、更快推理速度（33%加速）

Guo et al., Compressing Large Language Models by Joint Sparsification and Quantization, ICML 2024

扩散模型：时序特征保持量化

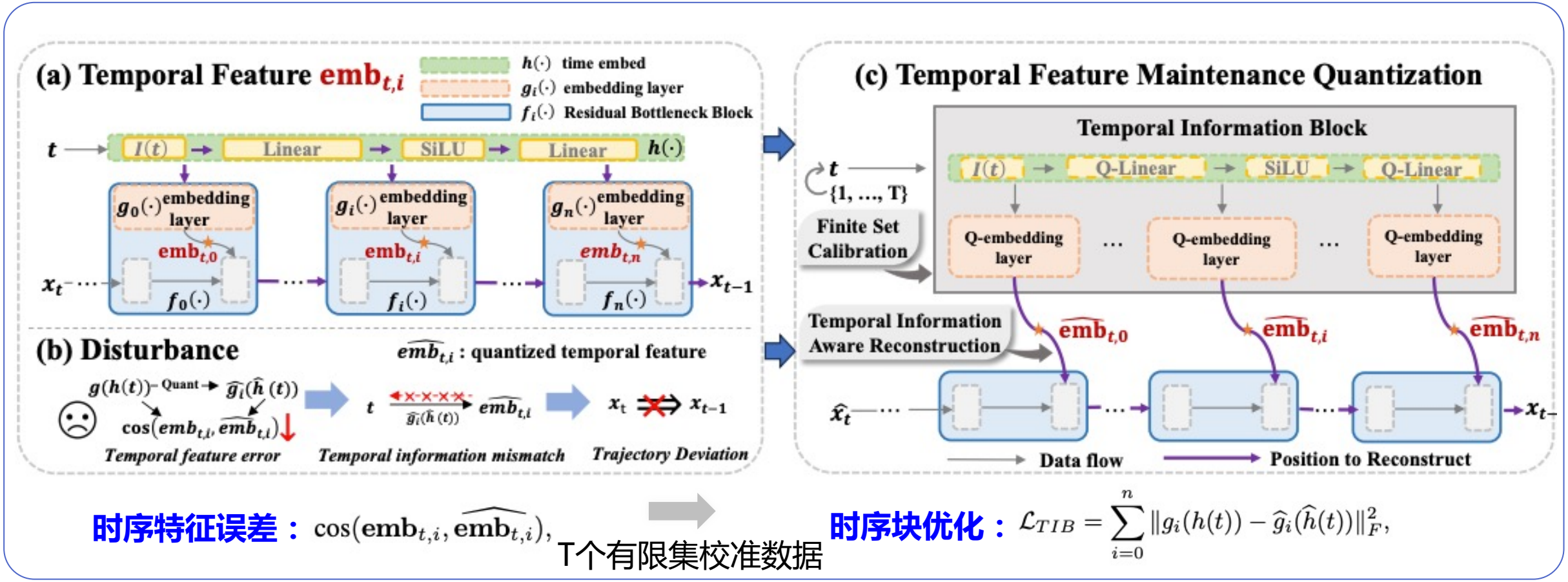
■ 问题：扩散模型中的时序特征对生成效果有较大影响



Huang et al., TFMQ-DM: Temporal Feature Maintenance Quantization for Diffusion Models. CVPR 2024 (highlight).

扩散模型：时序特征保持量化

■ **思路：提出时序特征保持思想，时序块优化+有限集校准实现精准量化**



时序特征误差： $\cos(\mathrm{emb}_{t,i}, \widehat{\mathrm{emb}}_{t,i})$,

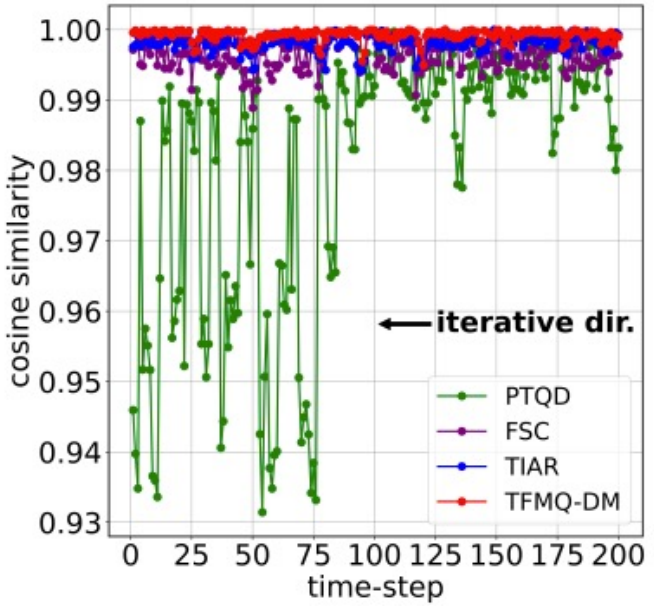
T个有限集校准数据

时序块优化： $\mathcal{L}_{TIB} = \sum_{i=0}^{n} \|g_i(h(t)) - \widehat{g}_i(\hat{h}(t))\|_F^2,$

扩散模型：时序特征保持量化

■ 扩散模型量化效果提升6%

Methods	Bits (W/A)	LSUN-Bedrooms 256 × 256		LSUN-Churches 256 × 256		CelebA-HQ 256 × 256		FFHQ 256 × 256	
		FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓	FID↓	sFID↓
Full Prec.	32/32	2.98	7.09	4.12	10.89	8.74	10.16	9.36	8.67
PTQ4DM* [42]	4/32	4.83	7.94	4.92	13.94	13.67	14.72	11.74	12.18
Q-Diffusion† [23]	4/32	4.20	7.66	4.55	11.90	11.09	12.00	11.60	10.30
PTQD* [10]	4/32	4.42	7.88	4.67	13.68	11.06	12.21	12.01	11.12
TFMQ-DM (Ours)	4/32	**3.60** (-0.60)	**7.61** (-0.05)	**4.07** (-0.48)	**11.41** (-0.49)	**8.74** (-2.32)	**10.18** (-1.82)	**9.89** (-1.71)	**9.06** (-1.24)
PTQ4DM* [42]	8/8	4.75	9.59	4.80	13.48	14.42	15.06	10.73	11.65
Q-Diffusion† [23]	8/8	4.51	8.17	4.41	12.23	12.85	14.16	10.87	10.01
PTQD [10]	8/8	3.75	9.89	4.89*	14.89*	12.76*	13.54*	10.69*	10.97*
TFMQ-DM (Ours)	8/8	**3.14** (-0.61)	**7.26** (-0.91)	**4.01** (-0.40)	**10.98** (-1.25)	**8.71** (-4.05)	**10.20** (-3.34)	**9.46** (-1.23)	**8.73** (-1.28)
PTQ4DM [42]	4/8	20.72	54.30	4.97*	14.87*	17.08*	17.48*	11.83*	12.91*
Q-Diffusion† [23]	4/8	6.40	17.93	4.66	13.94	15.55	16.86	11.45	11.15
PTQD [10]	4/8	5.94	15.16	5.10*	13.23*	15.47*	17.38*	11.42*	11.43*
TFMQ-DM (Ours)	4/8	**3.68** (-2.26)	**7.65** (-7.51)	**4.14** (-0.52)	**11.46** (-1.77)	**8.76** (-6.71)	**10.26** (-6.60)	**9.97** (-1.45)	**9.14** (-2.01)



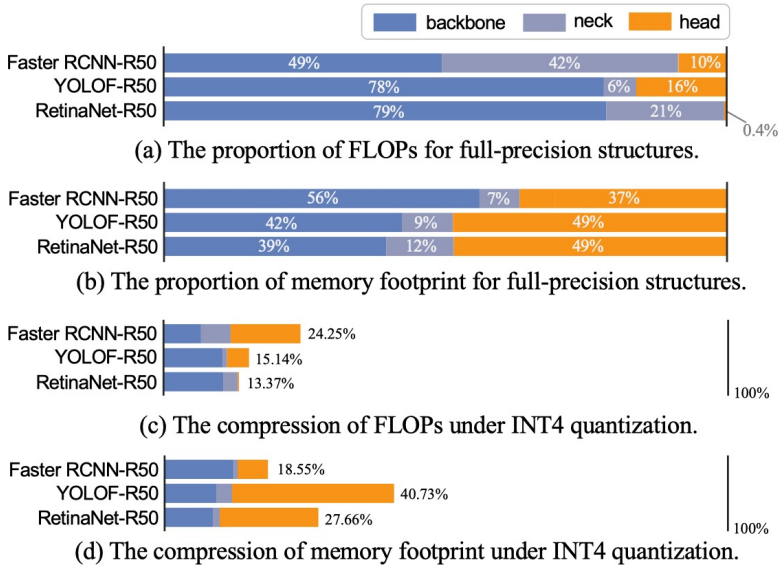
显著超越所有扩散模型量化方法，最多提升超6%，时序失配现象显著降低

Methods	Bits (W/A)	ImageNet 256 × 256		
		IS↑	FID↓	sFID↓
Full Prec.	32/32	235.64	10.91	7.67
PTQ4DM [42]	4/32	-	-	-
Q-Diffusion* [23]	4/32	213.56	11.87	8.76
PTQD† [10]	4/32	201.78	11.65	9.06
TFMQ-DM (Ours)	4/32	**223.81** (+10.25)	**10.50** (-1.15)	**7.98** (-0.78)
PTQ4DM [42]	8/8	161.75	12.59	-
Q-Diffusion* [23]	8/8	187.65	12.80	9.87
PTQD [10]	8/8	153.92	11.94	8.03
TFMQ-DM (Ours)	8/8	**198.86** (+11.21)	**10.79** (-1.15)	**7.65** (-0.38)
PTQ4DM [42]	4/8	-	-	-
Q-Diffusion* [23]	4/8	212.51	10.68	14.85
PTQD [10]	4/8	214.73	10.40	12.63
TFMQ-DM (Ours)	4/8	**221.82** (+7.09)	**10.29** (-0.11)	**7.35** (-5.28)

检测模型：回归结构友好量化

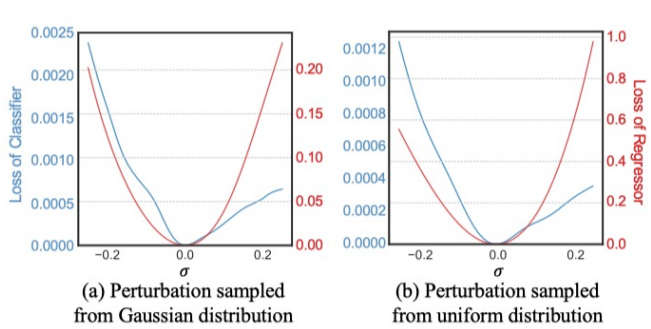
■ 问题：检测模型中的回归结构非均匀分布，更难找到合适量化参数

回归结构计算和体积占比高



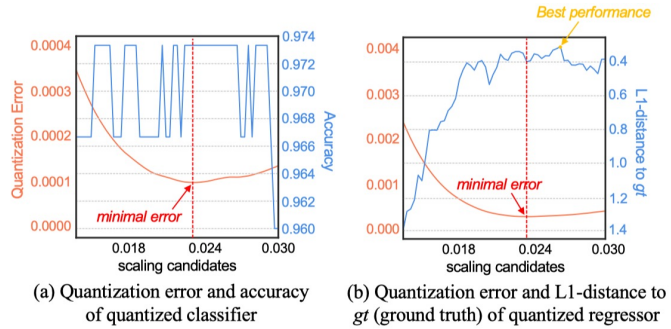
在单/双阶段检测架构中，回归结构的计算和参数量占比不可忽视

回归结构量化敏感



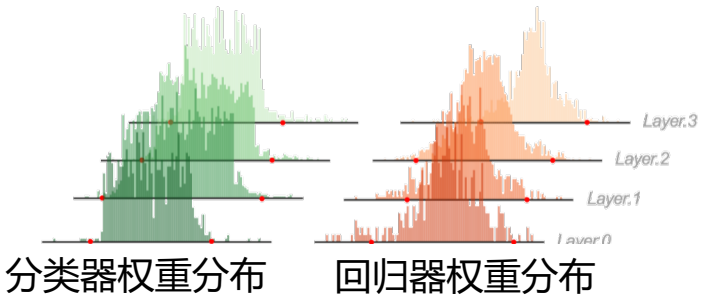
对回归和分类器分别叠加不同分布噪声，回归器输出误差更大

回归结构对量化扰动更敏感



最小化局部误差得到的量化参数往往不是最优

回归模型对局部累积误差更敏感



均匀量化器假设参数呈近似均匀分布，而回归结构的权重呈现非均匀长尾分布

均匀量化器无法较好呈现回归器原始参数分布

Ding et al., Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized Object Detector. IEEE CVPR 2024.

检测模型：回归结构友好量化

- **思路：提出回归结构友好量化框架，理论分析并解决回归结构参数分布非均匀问题，减少量化后参数误差**

理论分析回归架构难量化本质

$$W^* = \arg\max_W (P(Y|X,W)) = \arg\min_W (\|f(X) - Y\|_p),$$

以距离度量为目标函数的回归任务的权重期望

$$P(W|X,Y) \approx \prod_i \frac{1}{\lambda_1} \exp\left(-\frac{\|f(X_i) - Y_i\|_p}{\lambda_2}\right).$$

推导得到回归模型参数分布的后验概率分布

$p = 1, \lambda_2 = 2\lambda_1$ 参数呈Laplace分布

$p = 2, \lambda_2 = \lambda_1^2/\pi,$ 参数呈Gaussian分布

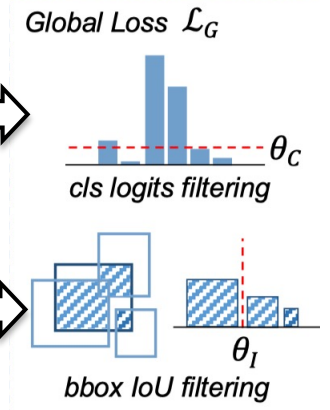
两步过滤的全局损失融合校准

$$b' = b \cdot \mathcal{I}_{HC} \cdot \mathcal{I}_{HI},$$

$$\mathcal{I}_{HC} = \begin{cases} 1, & \text{if } y \geq \theta_C \\ 0, & \text{otherwise,} \end{cases}$$

$$\hat{b}' = \hat{b} \cdot \mathcal{I}_{HC} \cdot \mathcal{I}_{HI},$$

$$\mathcal{I}_{HI} = \begin{cases} 1, & \text{if IoU}(b, \hat{b}) \geq \theta_I \\ 0, & \text{otherwise.} \end{cases}$$



分别过滤分类、回归置信度低的输出，获得准确的校准信息

$$\mathcal{L}_G = \frac{1}{n}\sum_{i=1}^{n}\left(L_{CE}(y_i, \hat{y}_i) + \lambda L_p(b_i', \hat{b}_i')\right)$$

两类损失联合对回归结构进行量化校准

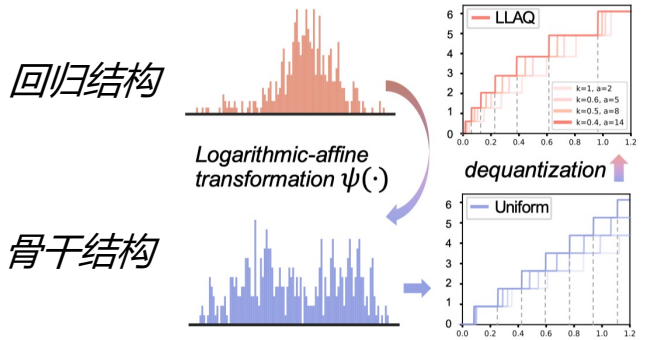
可学习对数仿射量化器

$$\psi(x) = k^* \log x + a^*$$

以零点为界，分段将参数映射到对数空间上，符合参数类Laplace分布形状

$$\psi(f(x|\mu,\lambda)) = \begin{cases} \frac{k^+}{\lambda}(\mu - x) - k^+ \log 2\lambda + a^+ & \text{if } x \geq \mu, \\ \frac{k^-}{\lambda}(x - \mu) - k^- \log 2\lambda + a^- & \text{otherwise.} \end{cases}$$

设置可学习参数逐层校准，拟合原始参数分布



Ding et al., Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized Object Detector. IEEE CVPR 2024.

检测模型：回归结构友好量化

■ 检测准确率超同类方法，更少漏检、更高精度，INT4 接近全精度

Method	#Bit(W/A)	RetinaNet		YOLOF	Faster RCNN		Mask RCNN	
		ResNet-50	ResNet-101	ResNet-50	ResNet-50	ResNet-101	ResNet-50	ResNet-101
Full-precision	32/32	37.4	38.9	37.5	38.5	39.8	39.2	40.8
BRECQ	2/4	14.0	18.7	10.8	12.5	13.0	11.0	12.0
PD-Quant	2/4	19.3	20.6	15.4	1.7	6.1	11.8	10.8
QDrop	2/4	19.9	22.9	17.4	17.8	19.9	18.2	19.9
Reg-PTQ (Ours)	**2/4**	**23.9**	**24.8**	**19.3**	**19.1**	**21.5**	**19.1**	**20.7**
AdaRound	3/3	19.3	20.7	7.7	21.2	22.8	21.6	22.6
AdaQuant	3/3	21.1	19.9	13.3	4.8	5.8	4.5	4.4
BRECQ	3/3	22.8	24.6	18.4	16.7	16.5	15.9	15.2
PD-Quant	3/3	24.5	25.6	22.2	14.0	14.0	18.7	17.3
QDrop	3/3	26.5	26.8	25.8	23.6	24.1	24.4	24.7
Reg-PTQ (Ours)	**3/3**	**28.1**	**28.3**	**27.3**	**28.1**	**29.1**	**28.4**	**28.8**
AdaRound	4/4	20.5	20.8	17.1	0.6	23.8	24.3	24.8
AdaQuant	4/4	33.5	34.5	25.6	12.8	14.5	12.0	14.6
BRECQ	4/4	34.2	35.8	29.0	28.8	30.8	31.7	30.1
PD-Quant	4/4	33.2	33.4	31.4	25.7	28.3	27.6	27.5
QDrop	4/4	34.1	35.1	33.4	33.7	34.4	34.5	35.6
SubSetQ	4/4	33.4	35.0	31.8	33.3	35.4	34.9	36.8
Reg-PTQ (Ours)	**4/4**	**36.7**	**35.9**	**34.3**	**36.7**	**36.2**	**36.4**	**37.2**

极低比特下优势更为明显，远超同类量化方法

目标检测任务
更少漏检率

FP　　QDrop　　Ours

#Bit(W/A)	Quantize Backbone & Neck		Fully Quantize	
	FLOPs (G)	Storage (M)	FLOPs (G)	Storage (M)
2/4	25.48	21.78	12.14	5.97
4/4	35.24	23.46	22.65	8.70
4/8	54.75	23.46	43.95	8.70

(a) Faster RCNN ResNet-50. The full-precision one has 171.8 GFLOPs and 46.91 M Storage while processing one sample.

#Bit(W/A)	Quantize Backbone & Neck		Fully Quantize	
	FLOPs (G)	# Storage (M)	FLOPs (G)	Storage (M)
2/4	8.06	37.11	7.89	14.9
4/4	14.73	39.08	14.46	18.42
4/8	28.07	39.08	27.84	18.42

(b) RetinaNet ResNet-50. The full-precision one has 108.10 GFLOPs and 66.60 M Storage while processing one sample.

不同比特下单双阶段检测模型理论计算和参数量

DataType	Latency(ms)	Storage(MB)
Float32	796.4	129.7
INT16	438.4	68.6
INT4*	132.8	38.6
INT4	84.5	22.8

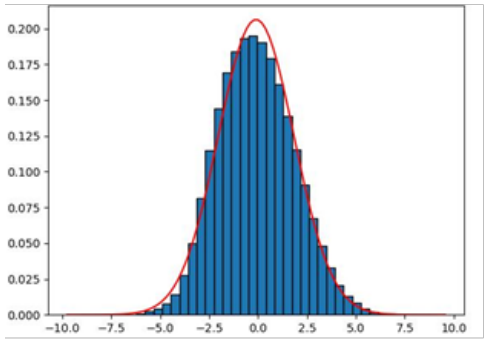
TVM部署实测
推理延迟

量化回归结构对检测模型
加快推理、减少存储收益显著

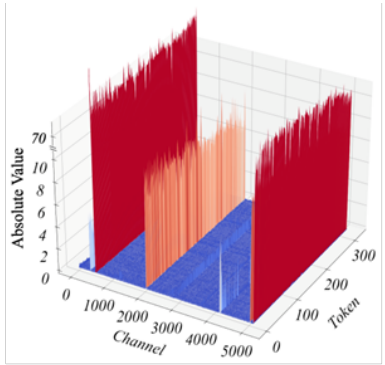
Ding et al., Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized Object Detector. IEEE CVPR 2024.

分割模型：SAM量化

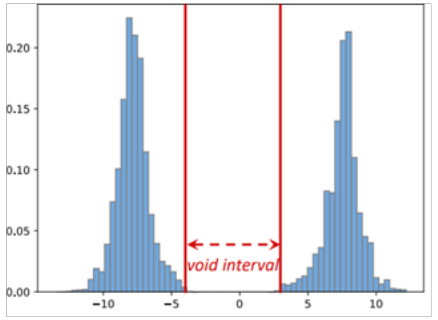
■ 问题：现有方法对双峰分布及多样的后软最大值分布信息损失严重



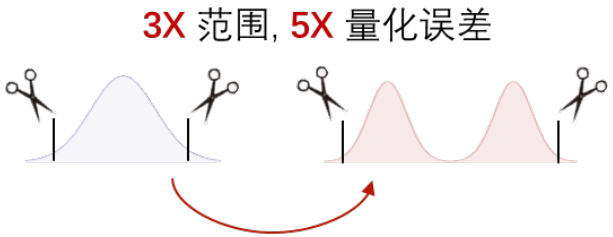
卷积神经网络



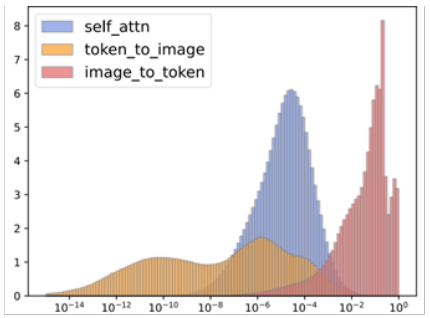
大语言模型

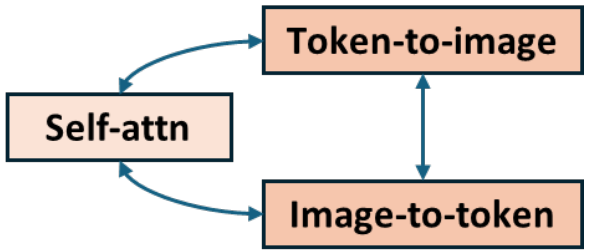


SAM模型



量化瓶颈





已有SOTA工作忽视双峰分布

逐张量、逐通道分析其性质

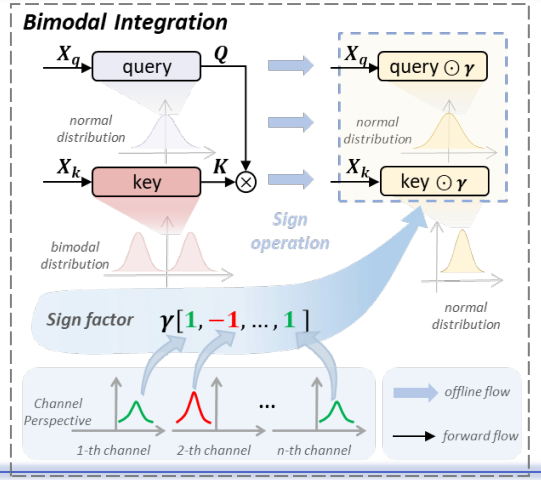
利用等价矩阵乘，离线迁移双峰分布

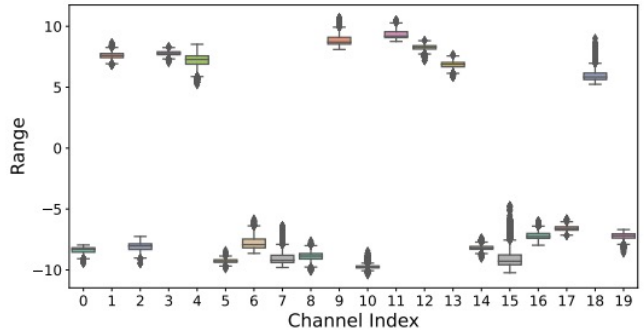
Lv et al., PTQ4SAM: Post-Training Quantization for Segment Anything, CVPR 2024

分割模型：SAM量化

■ 思路：提出双峰迁移和自适应粒度量化

双峰整合





双峰分布逐通道视角：单峰

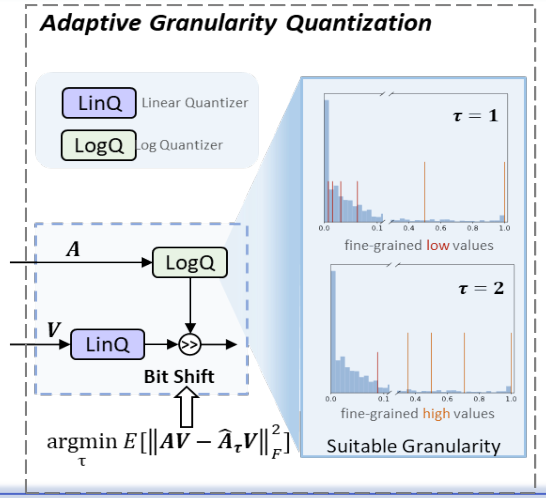
$$QK^\top = \underbrace{(X_qW_q + b_q)}_{\text{normal distribution}}\underbrace{(X_kW_k + b_k)^\top}_{\text{bimodal distribution}},$$

$$\gamma_j = \begin{cases} +1, & \text{if } \mathtt{mean}(K_{:,j}) \geq 0 \\ -1, & \text{otherwise} \end{cases},$$

$$QK^\top = ((X_qW_q + b_q) \odot \gamma)((X_kW_k + b_k)^\top \odot \gamma^\top)$$
$$= \underbrace{(X_qW_q' + b_q')}_{\text{normal distribution}}\underbrace{(X_kW_k' + b_k')^\top}_{\text{normal distribution}}. \quad (7)$$

离线转换：query和key矩阵乘等价性

自适应粒度量化



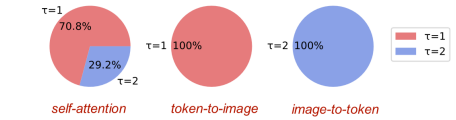
$$a_q = \mathtt{clamp}(\lfloor -\log_{2^{\frac{1}{\tau}}} \frac{a}{s} \rceil, 0, 2^k - 1), \qquad \hat{a} = s_a \cdot 2^{-\frac{a_q}{\tau}}.$$

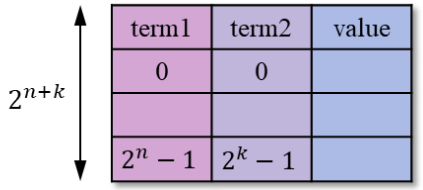
$$\hat{a} \cdot \hat{v} = s_a \cdot s_v \cdot LUT((-a_q)\%\tau, v_q >> \lceil \frac{a_q}{\tau} \rceil),$$

$$\arg\min_\tau \mathbb{E}\left[\|AV - \hat{A}_\tau V\|_F^2\right],$$

$$\hat{a} \cdot \hat{v} = s_a \cdot s_v \cdot 2^{\frac{(-a_q)\%\tau}{\tau}} \cdot 2^{\lfloor -\frac{a_q}{\tau} \rfloor} \cdot v_q$$
$$= s_a \cdot s_v \cdot \underbrace{2^{\frac{(-a_q)\%\tau}{\tau}}}_{(12-1)} \cdot v_q >> \lceil \frac{a_q}{\tau} \rceil,$$

重访对数量化器：硬件友好型证明





小型查找表实现高效乘加运算

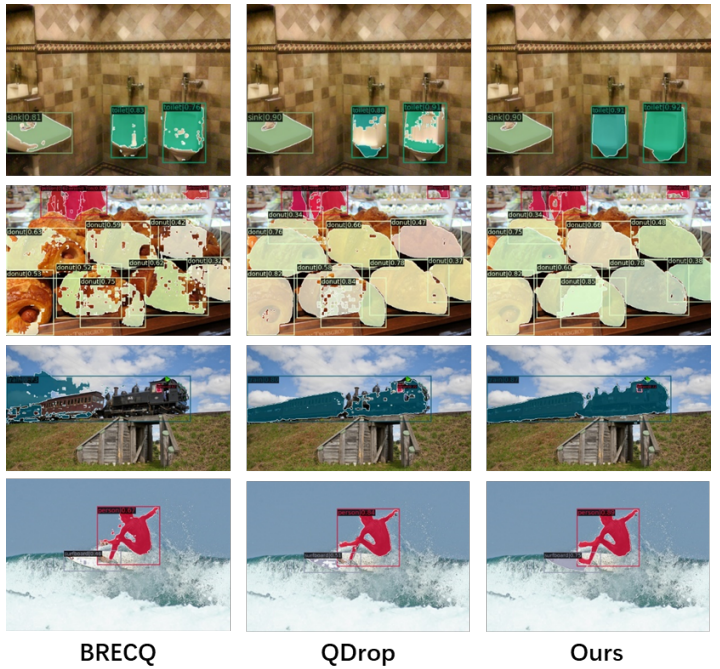
Lv et al., PTQ4SAM: Post-Training Quantization for Segment Anything, CVPR 2024

分割模型：SAM量化

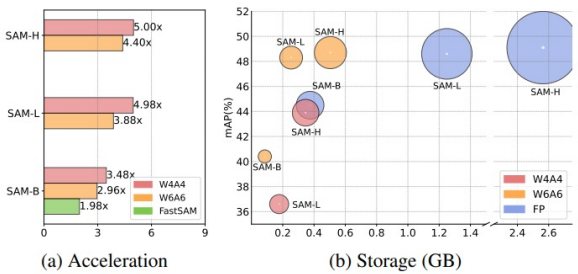
■ 在语义分割、实例分割、目标检测任务上远超目前SOTA方法

Detector	Methods	SAM-B			SAM-L			SAM-H		
		FP	W6A6	W4A4	FP	W6A6	W4A4	FP	W6A6	W4A4
Faster R-CNN [45]	MinMax [17]	33.4	9.2	-	36.4	32.9	-	37.2	31.9	-
	Percentile [60]		10.9	-		33.5	-		32.0	-
	OMSE [5]		11.9	-		33.9	5.4		33.1	7.4
	PTQ4SAM-S		**15.4**	-		**35.7**	**18.1**		**36.0**	**24.1**
	AdaRound [42]		23.1	-		34.3	8.7		33.7	14.5
	BRECQ [26]		24.1	-		34.2	10.7		33.7	15.1
	QDrop [57]		29.3	13.0		35.2	22.6		36.3	32.3
	PTQ4SAM-L		**30.3**	**16.0**		**35.8**	**28.7**		**36.5**	**33.5**
YOLOX [9]	MinMax [17]	37.0	10.7	-	40.4	37.5	-	41.0	36.1	-
	Percentile [60]		12.0	-		38.0	-		36.3	-
	OMSE [5]		13.5	-		38.4	6.1		37.5	7.8
	PTQ4SAM-S		**17.4**	-		**40.0**	**20.6**		**40.3**	**26.7**
	AdaRound [42]		26.4	-		38.9	11.1		38.3	16.7
	BRECQ [26]		26.1	-		38.9	12.0		38.3	16.3
	QDrop [57]		33.6	13.3		39.7	25.3		40.4	35.8
	PTQ4SAM-L		**34.3**	**18.4**		**40.3**	**31.6**		**40.7**	**37.6**

首次实现6-bit实例分割任务接近无损



4-bit实现3.48倍加速
近8倍存储压缩



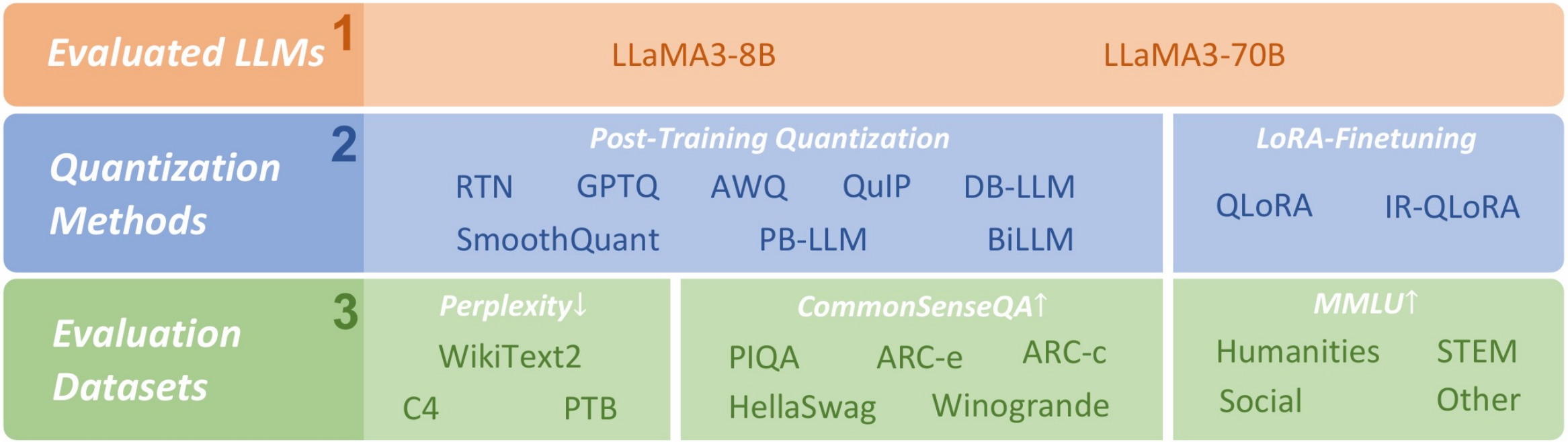
Base	Model	Methods	FP	W6A6	W4A4
31.78	+SAM-B	AdaRound [42]	33.15	32.34	31.78
		BRECQ [26]		32.27	31.78
		QDrop [57]		32.57	31.79
		PTQ4SAM-L		**32.65**	**31.85**
	+SAM-L	AdaRound [42]	33.61	32.99	31.97
		BRECQ [26]		33.04	31.98
		QDrop [57]		33.58	32.67
		PTQ4SAM-L		**33.66**	**32.82**
	+SAM-H	AdaRound [42]	33.63	33.49	32.17
		BRECQ [26]		33.46	32.12
		QDrop [57]		33.49	32.93
		PTQ4SAM-L		**33.66**	**33.10**

Model	Methods	FP	W6A6	W4A4
SAM-B	AdaRound [42]	64.1	34.05	-
	BRECQ [26]		34.40	-
	QDrop [57]		59.27	41.96
	PTQ4SAM-L		**60.33**	**44.18**
SAM-L	AdaRound [42]	64.2	63.44	23.18
	BRECQ [26]		63.60	26.89
	QDrop [57]		63.86	50.11
	PTQ4SAM-L		**63.91**	**56.29**
SAM-H	AdaRound [42]	64.6	62.73	24.45
	BRECQ [26]		62.58	25.98
	QDrop [57]		62.83	55.87
	PTQ4SAM-L		**64.36**	**56.01**

Lv et al., PTQ4SAM: Post-Training Quantization for Segment Anything, CVPR 2024

LLM量化评测

- 问题：随着LLaMA3的发布，量化压缩领域亟待对其进行评估
- 思路：全面测试多个典型方法，提供最新的开源大模型量化基准



测试全面涵盖了多个典型方法与指标

2大模型，2类压缩需求，10种量化方法，3类评测任务，12个数据集

Huang et al., How good are low-bit quantized llama3 models? an empirical study, arXiv

LLM量化评测

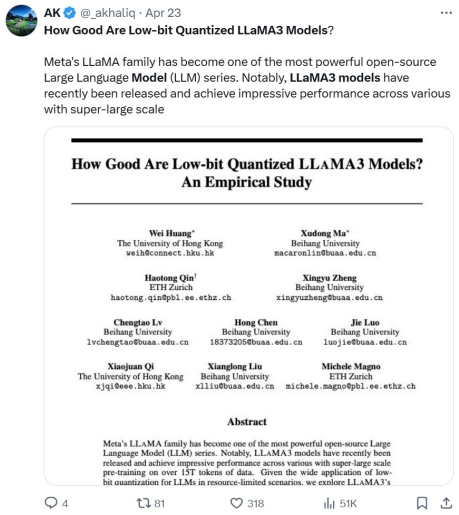
■ 高比特量化下性能依然优异，但随位宽降低性能下降显著

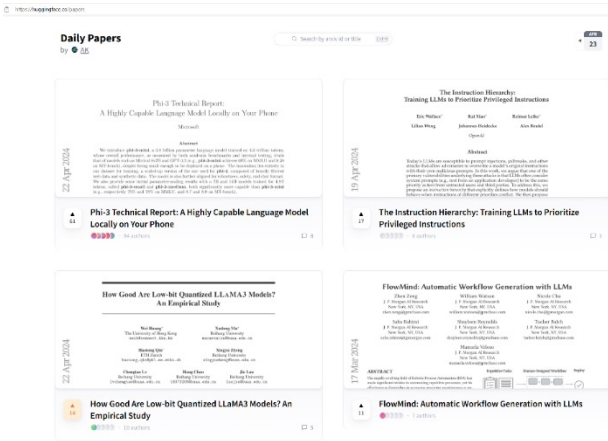
Method	#W	#A	#G	PPL↓			CommonSenseQA↑					
				WikiText2	C4	PTB	PIQA	ARC-e	ARC-c	HellaSwag	Wino	Avg.
LLaMA3	16	16	-	6.1	9.2	10.6	79.9	80.1	50.4	60.2	72.8	68.6
RTN	4	16	128	8.5	13.4	14.5	76.6	70.1	45.0	56.8	71.0	63.9
	3	16	128	27.9	1.1e2	95.6	62.3	32.1	22.5	29.1	54.7	40.2
	2	16	128	1.9E3	2.5E4	1.8E4	53.1	24.8	22.1	26.9	53.1	36.0
	8	16	-	6.2	9.5	11.2	79.7	80.8	50.4	60.1	73.4	68.9
	4	16	-	8.7	14.0	14.9	75.0	68.2	39.4	56.0	69.0	61.5
	3	16	-	2.2E3	5.6E2	2.0E3	56.2	31.1	20.0	27.5	53.1	35.6
	2	16	-	2.7E6	7.4E6	3.1E6	53.1	24.7	21.9	25.6	51.1	35.3
GPTQ	4	16	128	6.5	10.4	11.0	78.4	78.8	47.7	59.0	72.6	67.3
	3	16	128	8.2	13.7	15.2	74.9	70.5	37.7	54.3	71.1	61.7
	2	16	128	2.1E2	4.1E4	9.1E2	53.9	28.8	19.9	27.7	50.5	36.2
	8	16	-	6.1	9.4	10.6	79.8	80.1	50.2	60.2	72.8	68.6
	4	16	-	7.0	11.8	14.4	76.8	74.3	42.4	57.4	72.8	64.8
	3	16	-	13.0	45.9	37.0	60.8	38.8	22.3	41.8	60.9	44.9
	2	16	-	5.7E4	1.0E5	2.7E5	52.8	25.0	20.5	26.6	49.6	34.9
AWQ	4	16	128	6.6	9.4	11.1	79.1	79.7	49.3	59.1	74.0	68.2
	3	16	128	8.2	11.6	13.2	77.7	74.0	43.2	55.1	72.1	64.4
	2	16	128	1.7E6	2.1E6	1.8E6	52.4	24.2	21.5	25.6	50.7	34.9
	8	16	-	6.1	8.9	10.6	79.6	80.3	50.5	60.2	72.8	68.7
	4	16	-	7.1	10.1	11.8	78.3	77.6	48.3	58.6	72.5	67.0
	3	16	-	12.8	16.8	24.0	71.9	66.7	35.1	50.7	64.7	57.8
	2	16	-	8.2E5	8.1E5	9.0E5	55.2	25.2	21.3	25.4	50.4	35.5
QuIP	4	16	-	6.5	11.1	9.5	78.2	78.2	47.4	58.6	73.2	67.1
	3	16	-	7.5	11.3	12.6	76.8	72.9	41.0	55.4	72.5	63.7
	2	16	-	85.1	1.3E2	1.8E2	52.9	29.0	21.3	29.2	51.7	36.8
DB-LLM	2	16	128	13.6	19.2	23.8	68.9	59.1	28.2	42.1	60.4	51.8
PB-LLM	2	16	128	24.7	79.2	65.6	57.0	37.8	17.2	29.8	52.5	38.8
	1.7	16	128	41.8	2.6E2	1.2E2	52.5	31.7	17.5	27.7	50.4	36.0
BiLLM	1.1	16	128	28.3	2.9E2	94.7	56.1	36.0	17.7	28.9	51.0	37.9
SmoothQuant	8	8	-	6.3	9.2	10.8	79.5	79.7	49.0	60.0	73.2	68.3
	6	6	-	7.7	11.8	12.5	76.8	75.5	45.0	56.9	69.0	64.6
	4	4	-	4.3E3	4.0E3	3.6E3	54.6	26.3	20.0	26.4	50.3	35.5

PTQ方法下的LLaMA3-8B量化结果

Method	#W	MMLU↑					CommonSenseQA↑					
		Hums.	STEM	Social	Other	Avg.	PIQA	ARC-e	ARC-c	HellaSwag	Wino	Avg.
LLaMA3	16	59.0	55.3	76.0	71.5	64.8	79.9	80.1	50.4	60.2	72.8	68.6
NormalFloat	4	56.8	52.9	73.6	69.4	62.5	78.6	78.5	46.2	58.8	74.3	67.3
QLoRA	4	50.3	49.3	65.8	64.2	56.7	76.6	74.8	45.0	59.4	67.0	64.5
IR-QLoRA	4	52.2	49.0	66.5	63.1	57.2	76.3	74.3	45.3	59.1	69.5	64.9

PEFT方法下的LLaMA3-8B量化微调结果





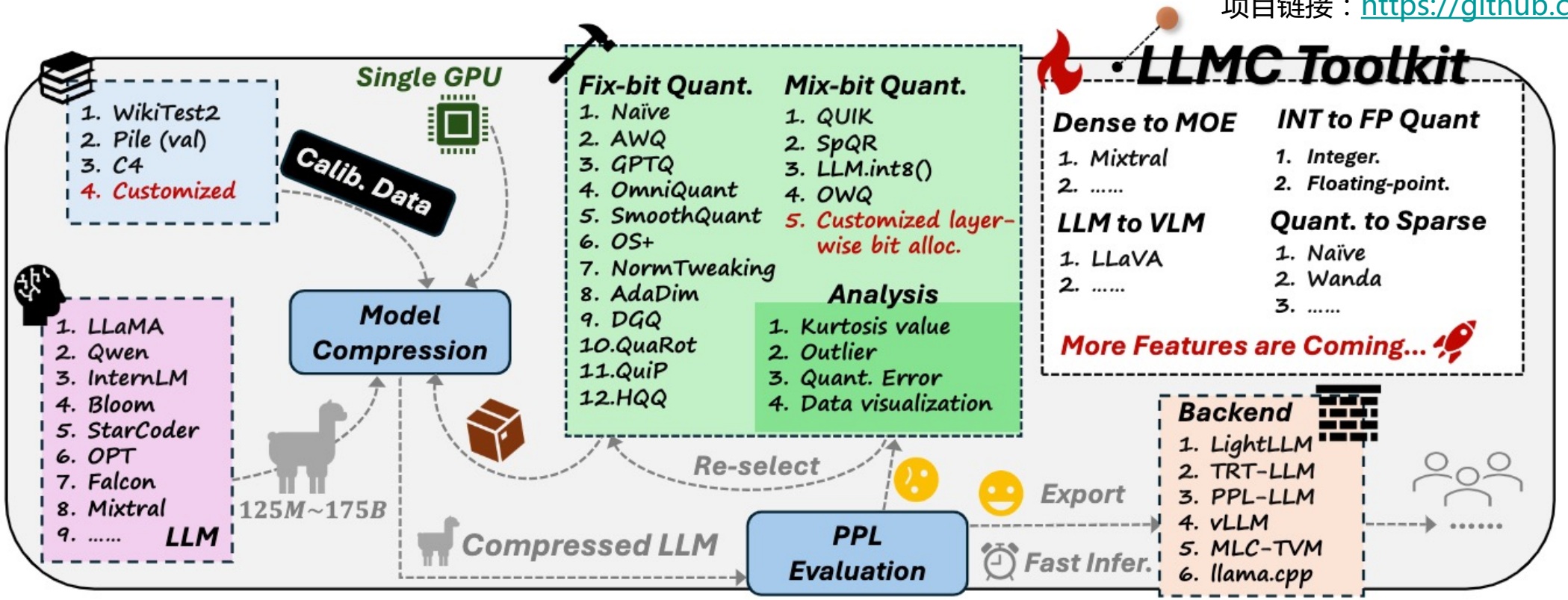
获各类平台媒体广泛关注

Huang et al., How good are low-bit quantized llama3 models? an empirical study, arXiv

大语言模型和多模态大模型压缩工具

- 问题：大语言模型压缩方法多样，模型种类繁多，成本高
- 思路：一套工具一站式解决大语言模型量化压缩

项目链接：https://github.com/ModelTC/llmc



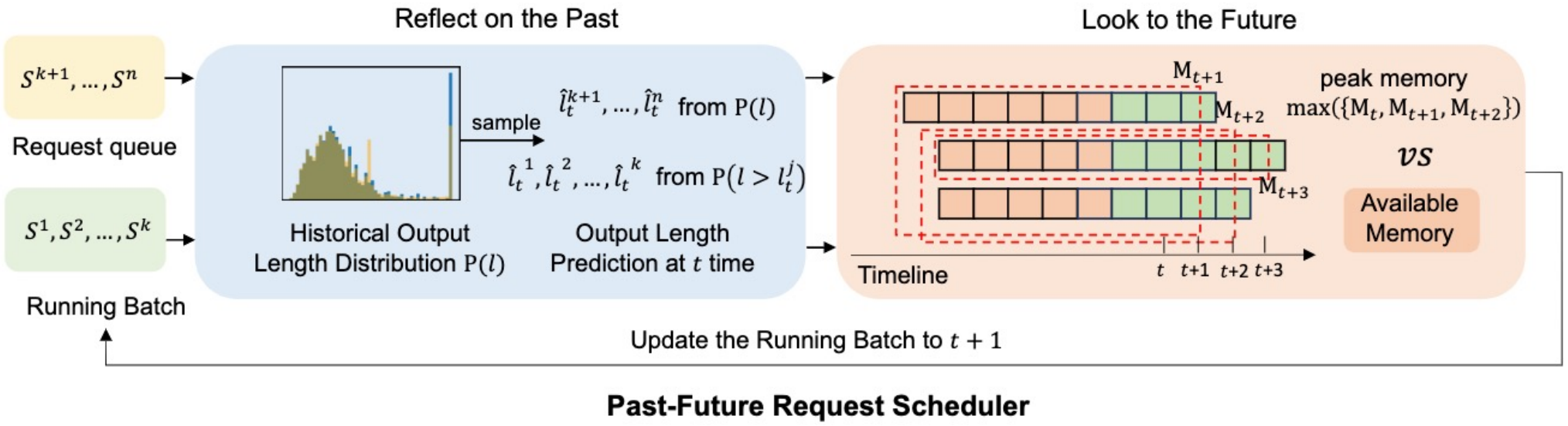
项目二维码

欢迎关注、使用和Star！

16种算法，12种模型，文本和视觉多种模态，6种硬件后端

多模态大模型高效推理服务系统 LightLLM

- **问题：大语言模型和多模态大模型服务输入输出动态变化大**
- **思路：一套集高效后端、极致压缩、精准调度于一体的大模型服务框架**



Past-Future Request Scheduler

项目二维码

欢迎关注、使用和Star！

被多个业界研究机构和公司关注、使用

Berkeley UNIVERSITY OF CALIFORNIA　香港科技大學　TRIPO　Tencent 腾讯　HUAWEI

多种模态，全面支持量化，细粒度显存管理和调度，极致吞吐性能

https://github.com/ModelTC/lightllm

目录

二值量化

■ 二值化（1-bit）：极限压缩位宽的量化，面向更高精度

- 硬件上利用专有的位运算指令加速计算
- 二值模型具有高达32x的压缩潜力

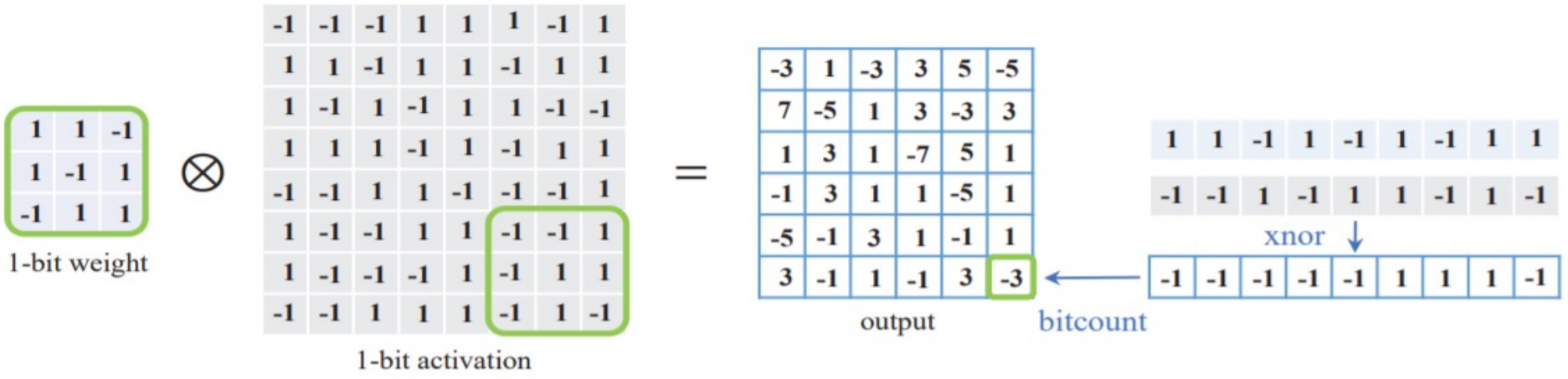
$$\mathbb{R} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit　　1-bit

$$\mathbf{B_x} = \mathtt{sign}(\mathbf{x}) = \begin{cases} +1, & if\ \mathbf{x} \geq 0 \\ -1, & otherwise. \end{cases}$$

$$Q_x(\mathbf{x}) = \alpha \mathbf{B_x},$$

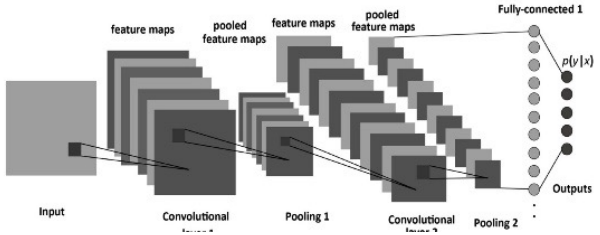
$$\mathbf{z} = \sigma(Q_w(\mathbf{w}) \otimes Q_a(\mathbf{a})) = \sigma(\alpha\beta(\mathbf{b_w} \odot \mathbf{b_a}))$$



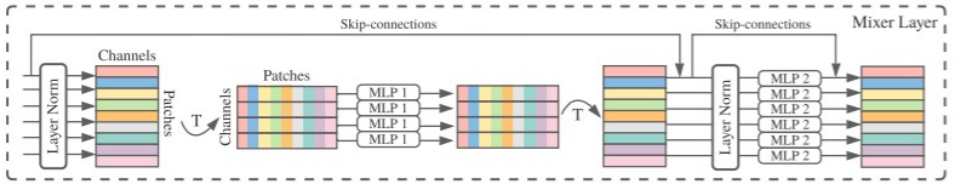
二值量化

■ 二值化（1-bit）：极限压缩位宽的量化，面向更高精度

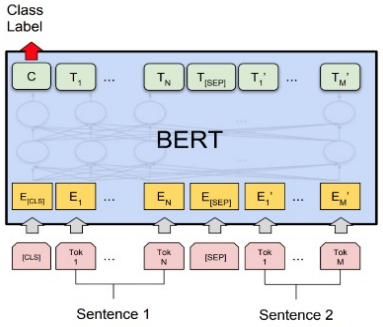
- **挑战**：精度损失大是阻碍二值化落地的最关键原因
- **实验观察**：**结构**是影响二值量化精度的主要因素之一



CNNs (ResNet, VGG, ...)



MLPs (PointNet, MLP-Mixer, ...)



Transformers (BERT, ViT,...)

典型架构如何实现高准确率的二值量化？

相关工作

■ 二值化（1-bit）：极限压缩位宽的量化，面向更高精度

- **提高二值量化模型精度**

XNOR-Net

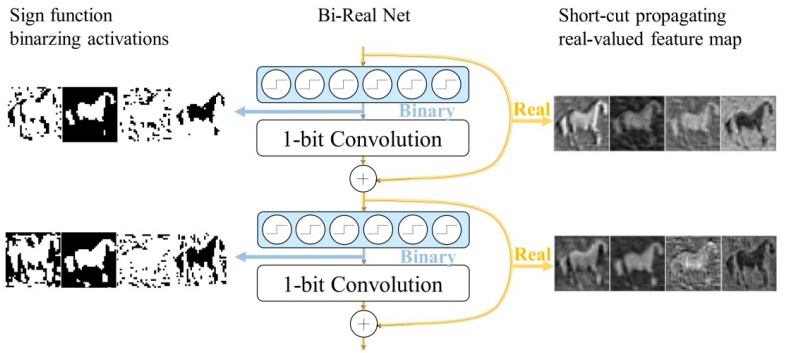
Mohammad Rastegari等人，华盛顿大学，ECCV 2016

通过最小化二值量化误差来获得精确的二值神经网络。

Bi-Real Net

Zechun Liu等人，香港科技大学，ECCV 2018

通过在ResNet中引入残差结构来获得更精确的二值神经网络。



- **将二值量化扩展至更多模型和任务**

BGNN

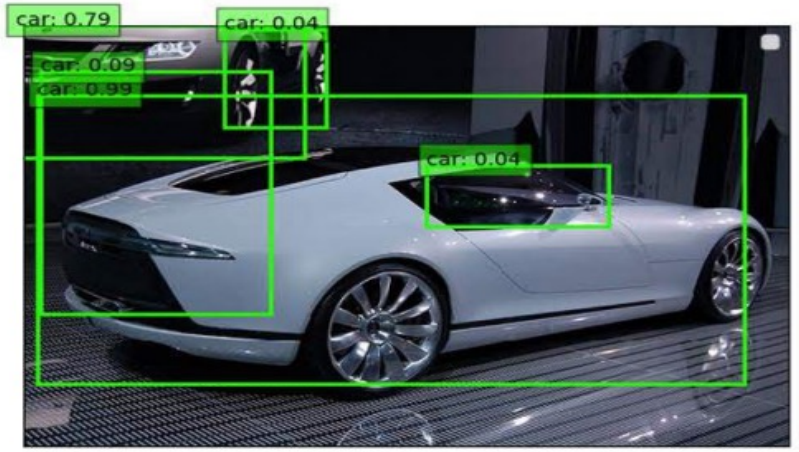
Mehdi Bahri等人，帝国理工大学，CVPR 2021

研究精确的二值图神经网络。

BiDet

Ziwei Wang等人，清华大学，CVPR 2020

研究高效的二值目标检测网络。



二值量化方法经典工作

经典二值量化方法总结

朴素二值量化方法

- Binary Connect
- Bitwise Neural Networks
- Binarized Neural Networks

$$\boldsymbol{I} * \boldsymbol{W} \approx sign(\boldsymbol{I}) \otimes sign(\boldsymbol{W})$$

减少量化误差的优化方法

$$J(\boldsymbol{b}, \alpha) = \|\boldsymbol{x} - \alpha \boldsymbol{b}\|^2$$
$$\alpha^*, \boldsymbol{b}^* = \underset{\alpha, \boldsymbol{b}}{\operatorname{argmin}} J(\boldsymbol{b}, \alpha)$$

- Binary Weight Networks
- XNOR-Net
- DoReFa-Net
- HORQ
- ABC-Net
- Two-Step Quantization
- LQ-Nets
- XNOR-Net++
- Real-to-Binary

改进网络损失的优化方法

$$\mathcal{L}^b_{total} = \mathcal{L}^b_{original} + \lambda \mathcal{L}^b_{Customized}$$

- Distilled Binary Neural Network
- Distillation and Quantization
- Apprentice
- Loss-Aware Binarization
- Incremental Network Quantization
- BNN-DL
- CI-BCNN
- Main/Subsidiary Network

减小梯度误差的优化方法

Customized
$ApproxFunc/QuantFunc/UpdateFunc$

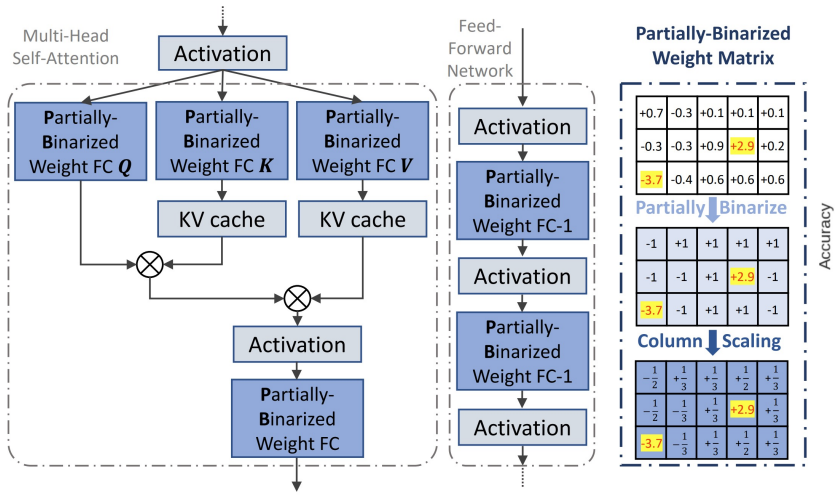
- Bi-Real Net
- Circulant Binary Convolutional Networks
- Half-wave Gaussian Quantization
- BNN+
- Differentiable Soft Quantization
- BCGD
- ProxQuant
- IR-Net

二值量化经典工作

二值化最新进展

大模型二值化现状：精度下降显著，校准/训练成本极高

PB-LLM (PTQ)

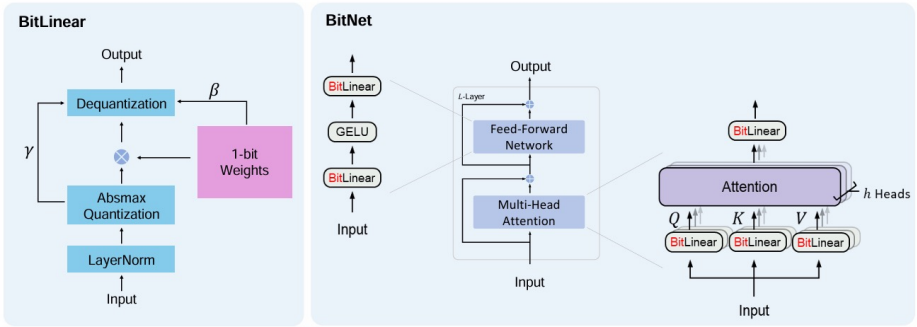


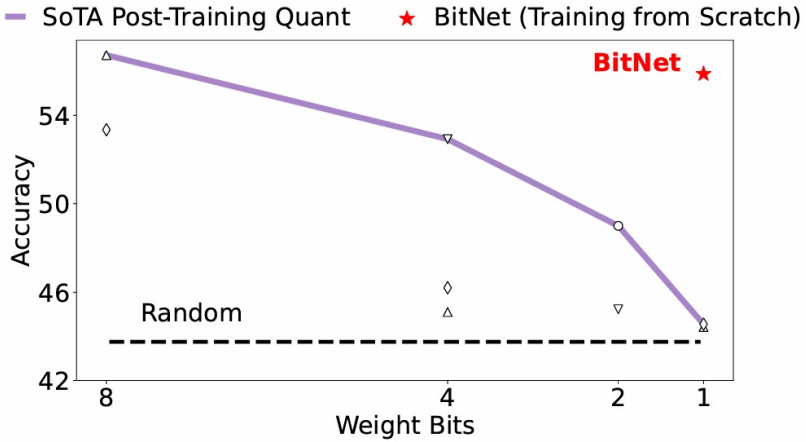
Method	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Avg
FP LLaMA-7B	76.8	79.3	76.1	70.0	73.0	48.0	57.6	68.7
RTN	71.2	77.3	72.7	66.9	68.8	46.4	52.8	65.2
SmoothQuant	67.7	76.0	69.4	66.7	66.9	43.0	50.6	63.0
LLM-QAT	75.5	78.3	74.0	69.0	70.0	45.0	55.4	66.6
PB-GPTQ 10%	62.3	55.9	27.7	49.3	29.3	20.1	10.6	36.5
PB-GPTQ 30%	73.5	74.9	47.5	64.9	61.3	32.4	25.2	54.2
PB-LLM 10%	68.9	67.8	68.1	67.4	58.7	42.9	50.6	60.6
PB-LLM 30%	75.7	78.0	74.3	69.7	69.0	45.6	55.8	66.9

部分显著权重仍保留高比特

低比特PTQ下与FP差距显著

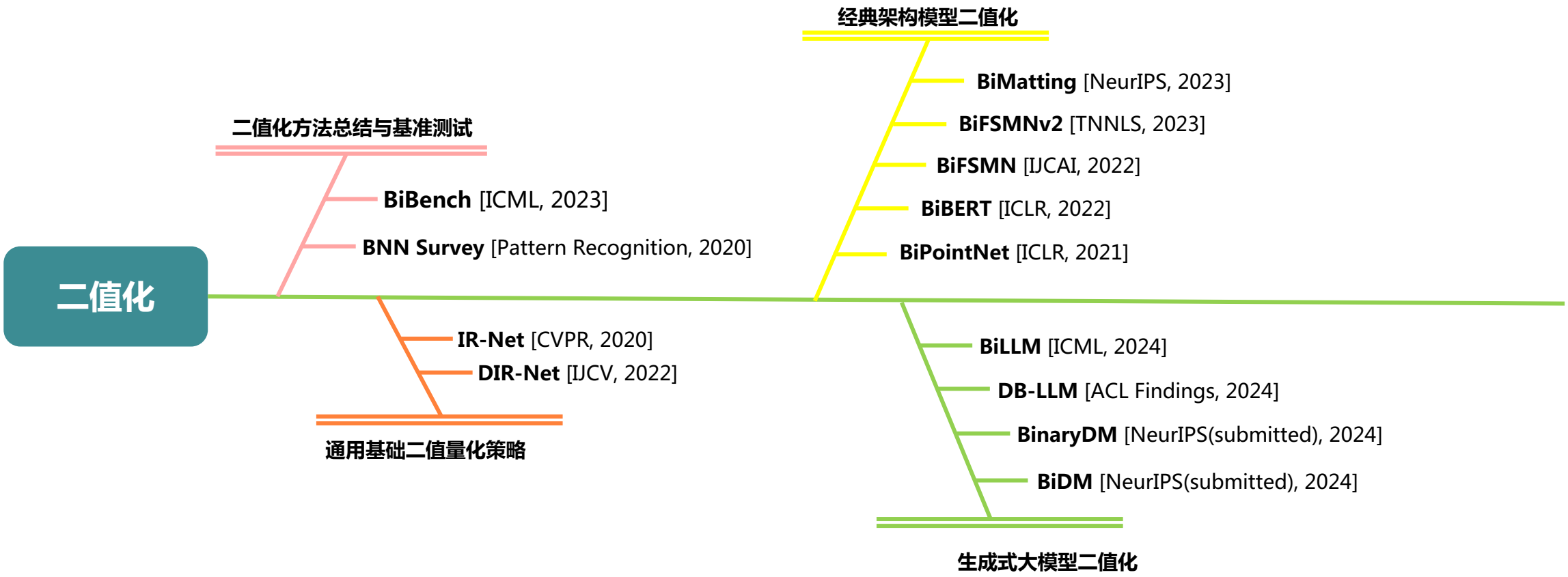
BitNet (QAT)





需要从头训练，资源开销大

相关工作


二值化
二值化方法总结与基准测试
BiBench [ICML, 2023]
BNN Survey [Pattern Recognition, 2020]
通用基础二值量化策略
IR-Net [CVPR, 2020]
DIR-Net [IJCV, 2022]
经典架构模型二值化
BiMatting [NeurIPS, 2023]
BiFSMNv2 [TNNLS, 2023]
BiFSMN [IJCAI, 2022]
BiBERT [ICLR, 2022]
BiPointNet [ICLR, 2021]
生成式大模型二值化
BiLLM [ICML, 2024]
DB-LLM [ACL Findings, 2024]
BinaryDM [NeurIPS(submitted), 2024]
BiDM [NeurIPS(submitted), 2024]

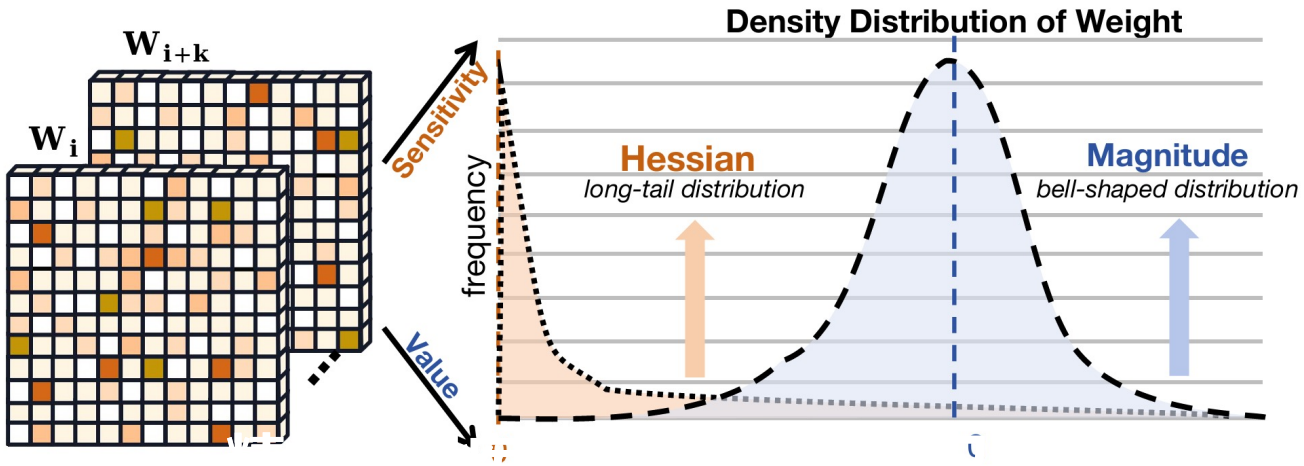
LLM二值化：1 + 0.08

■ 问题：现有的LLM 的 PTQ 方法在 2 bit 以下性能崩溃

PPL↓

Model	Method	Block Size	Weight Bits	7B	13B	30B	65B/70B*
LLaMA	Full Precision	-	16.00	5.68	5.09	4.10	3.53
	RTN	-	2.00	106767.34	57409.93	26704.36	19832.87
	GPTQ	128	2.00	152.31	20.44	13.01	8.78
	RTN	-	1.00	168388.00	1412020.25	14681.76	65253.24
	GPTQ	128	1.00	267001.72	113894.12	67093.73	25082.88
	PB-LLM †	128	1.70	102.36	36.60	33.67	12.53

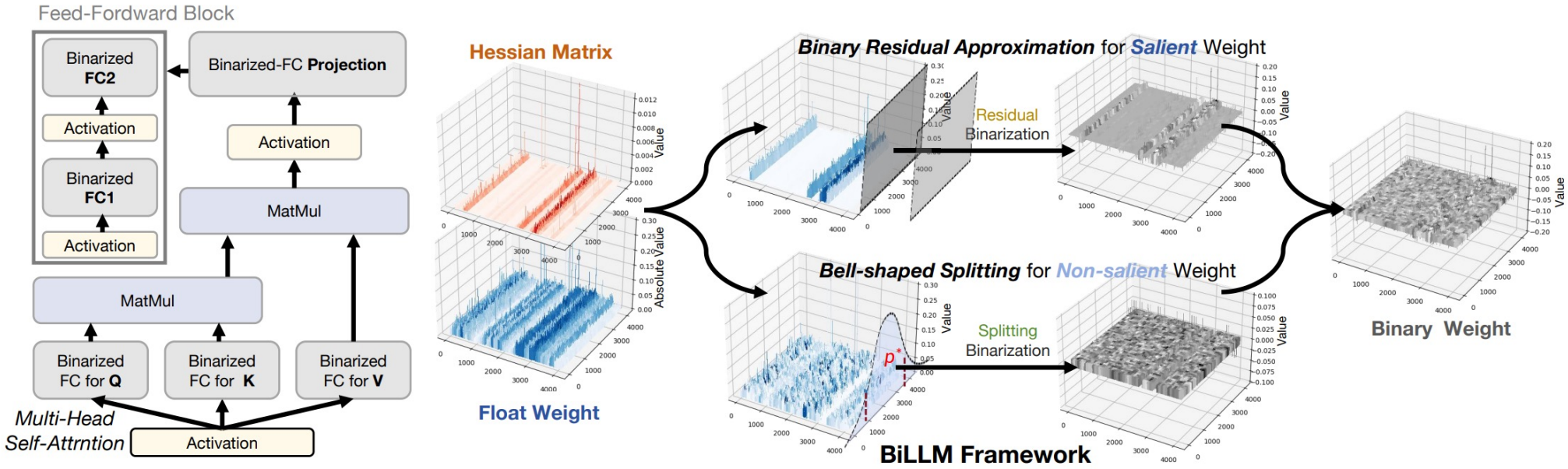
现有的以GPTQ为代表的
PTQ方法在2bit与1bit下严重崩溃



通过Hessian度量的权重(sensitivity)
与通过Magnitude度量的权重(value)
分别呈现了长尾分布与钟形分布

LLM二值化：1 + 0.08

- **思路：在基于Hessian度量的指导下选择按列分割的显著权重，再进行残差近似对保持钟形分布的剩余权重，采用细致的搜索进行断点分割**



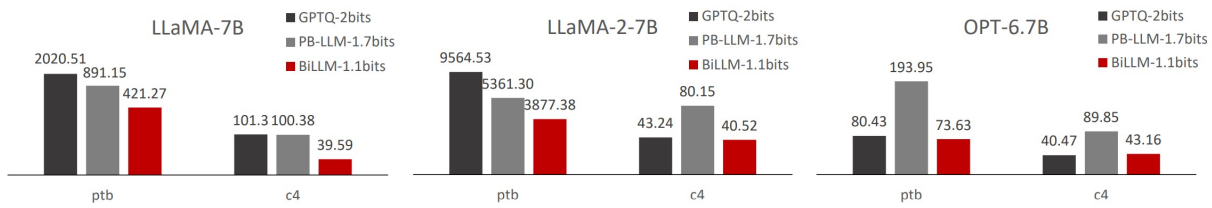
重要性度量： $s_i = \frac{w_i^2}{[\mathbf{H}^{-1}]_{ii}^2},$ **显著列选择：** $\arg\min_{\mathbf{W}_{\text{uns}}} ||\mathbf{W} - (\alpha_{\text{sal}}\,\text{sign}(\mathbf{W}_{\text{sal}}) \cup \alpha_{\text{uns}}\,\text{sign}(\mathbf{W}_{\text{uns}}))||^2,$

残差二值化： $\alpha_o^*, \mathbf{B}_o^* = \arg\min_{\alpha_o, \mathbf{B}_o} ||\mathbf{W} - \alpha_o \mathbf{B}_o||^2), \;\; \alpha_r^*, \mathbf{B}_r^* = \arg\min_{\alpha_r, \mathbf{B}_r} ||(\mathbf{W} - \alpha_o^* \mathbf{B}_o^*) - \alpha_r \mathbf{B}_r||^2),$

二值误差： $\theta_{q,p}^2 = ||\mathbf{W}_s - \alpha_s \mathbf{B}_s||^2 + ||\mathbf{W}_c - \alpha_c \mathbf{B}_c||^2,$ **优化目标：** $p^* = \arg\min_p (\theta_{q,p}^2).$

LLM二值化：1 + 0.08

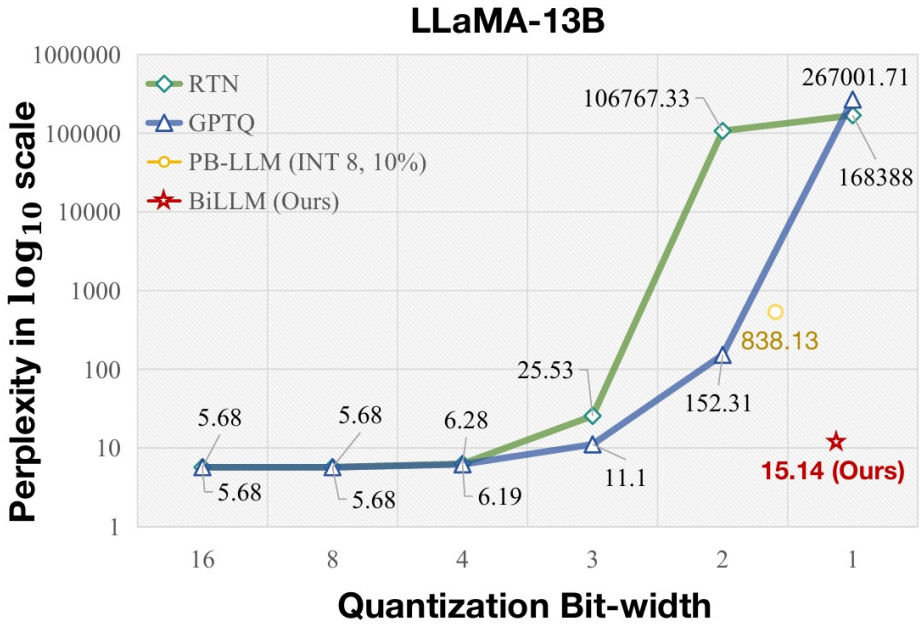
■ 1.08-bits的BiLLM首次实现了超低比特模型的高精度推理



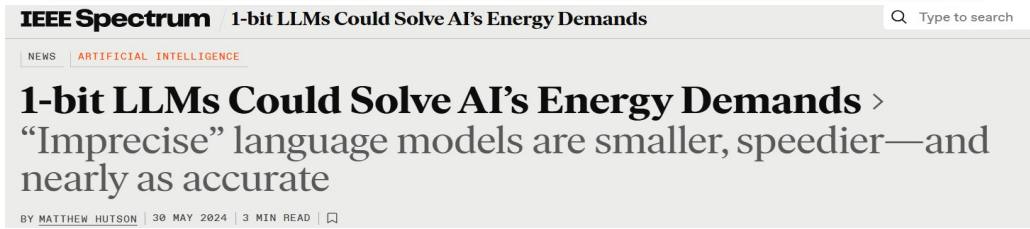
BiLLM在各种模型与数据集上表现出了性能优势

Model	Method	Block Size	Weight Bits	7B	13B	30B	65B/70B*
LLaMA	Full Precision	-	16.00	5.68	5.09	4.10	3.53
	RTN	-	2.00	106767.34	57409.93	26704.36	19832.87
	GPTQ	128	2.00	152.31	20.44	13.01	8.78
	RTN	-	1.00	168388.00	1412020.25	14681.76	65253.24
	GPTQ	128	1.00	267001.72	113894.12	67093.73	25082.88
	PB-LLM †	128	1.70	102.36	36.60	33.67	12.53
	BiLLM ‡	128	**1.09**	**35.04**	**15.14**	**10.52**	**8.49**
LLaMA2	Full Precision	-	16.00	5.47	4.88	N/A	3.32
	RTN	-	2.00	17788.93	51145.61	N/A	26066.13
	GPTQ	128	2.00	60.45	19.70	N/A	9.12
	RTN	-	1.00	157058.34	47902.32	N/A	160389.91
	GPTQ	128	1.00	115905.67	9387.80	N/A	74395.42
	PB-LLM †	128	1.70	69.20	151.09	N/A	28.37
	BiLLM ‡	128	**1.08**	**32.48**	**16.77**	N/A	**8.41**

LLaMA模型家族与各方法在wikitext2上的PPL



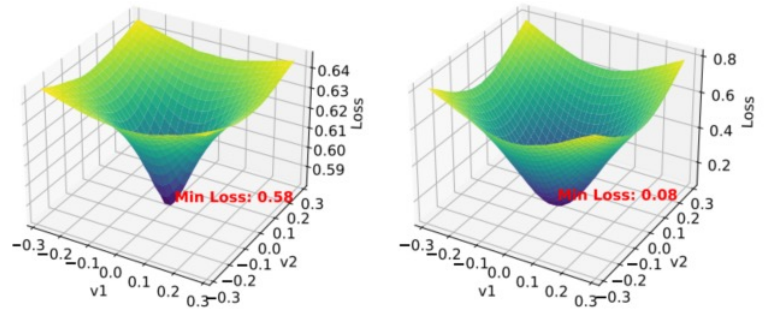
BiLLM在LLaMA-13B的二值化下表现出优异的性能，超越了GPTQ、PB-LLM等先进PTQ量化方法



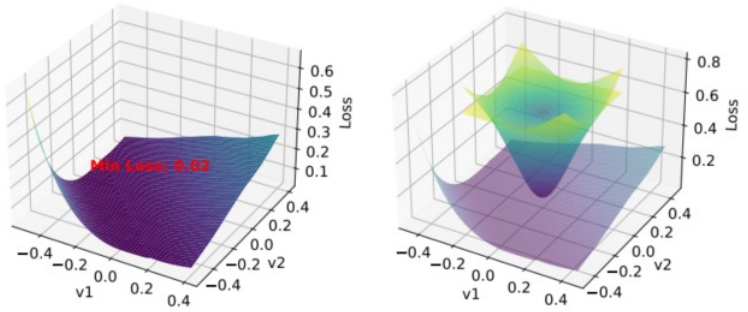
双重二值化：1 + 1

■ 问题：提高大模型在极低比特量化下的精度

极低比特量化下权重优化平面陡峭

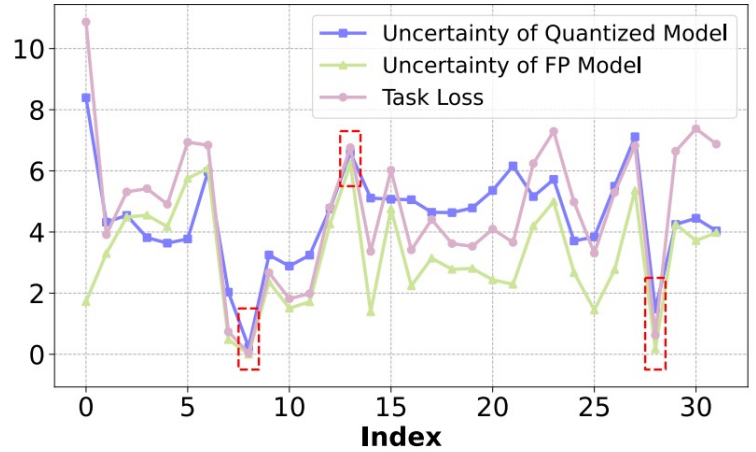


(a) Binarization. (b) 2-bit quantization.



(c) FDB. (d) All Together.

量化模型具有更高的不确定性



$$\mathcal{H}(\boldsymbol{P}) = -\sum_{i=1}^{C} p_i \log(p_i),$$

使用信息熵衡量不确定度

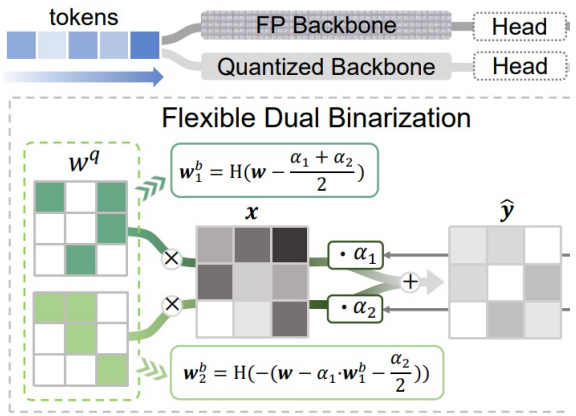
需恢复模型在极低比特量化下的表达能力

Chen et al., DB-LLM: Accurate Dual-Binarization for Efficient LLMs, ACL Findings 2024

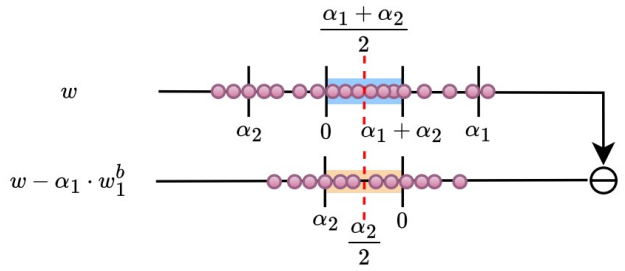
双重二值化：1 + 1

- **思路：2-bit量化权重分成两个独立的二值集合并引入偏差感知蒸馏，引导量化模型更关注模糊样本，细化量化参数提升性能**

双重二值稀疏权重



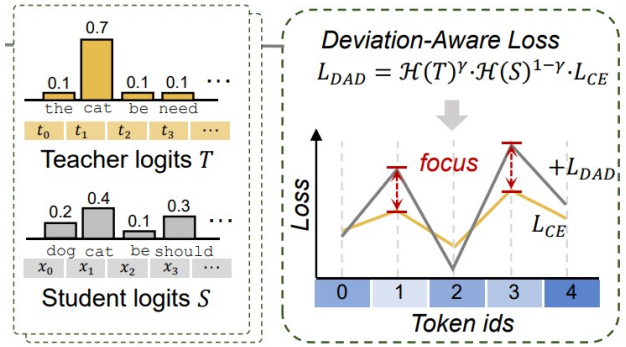
$$\hat{\boldsymbol{w}} = s \cdot \boldsymbol{w}^q = \alpha_1 \cdot \boldsymbol{w}_1^b + \alpha_2 \cdot \boldsymbol{w}_2^b$$
$$\alpha_1 := 2s, \alpha_2 := -s$$

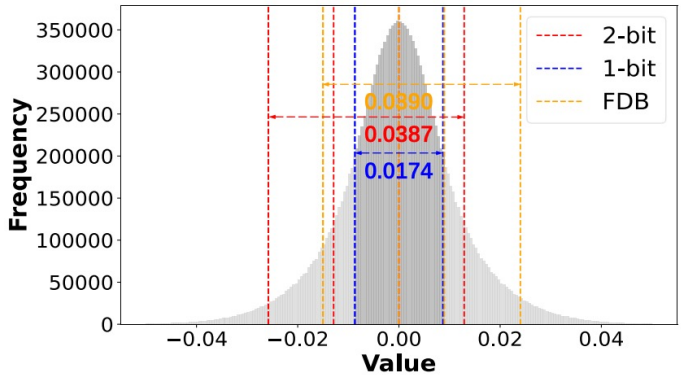


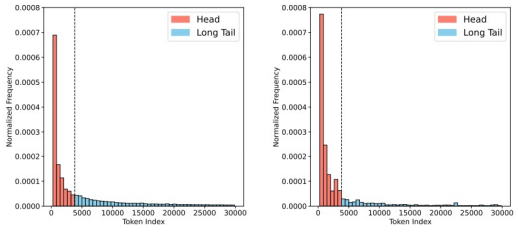
$$\boldsymbol{w}_1^b = \mathrm{H}(\boldsymbol{w} - \frac{\alpha_1+\alpha_2}{2}),$$
$$\boldsymbol{w}_2^b = \mathrm{H}(-(\boldsymbol{w} - \alpha_1 \cdot \boldsymbol{w}_1^b - \frac{\alpha_2}{2})),$$
$$\hat{\boldsymbol{y}} = \alpha_1 \cdot (\boldsymbol{w}_1^b \otimes \boldsymbol{x}) + \alpha_2 \cdot (\boldsymbol{w}_2^b \otimes \boldsymbol{x}),$$

将2-bit权重分割为两个独立的1-bit权重分别进行二值矩阵乘法

偏差感知蒸馏







(a) FP LLaMA-1-7B. (b) 2-bit LLaMA-1-7B.

引导模型专注于模糊难样本

Chen et al., DB-LLM: Accurate Dual-Binarization for Efficient LLMs, ACL Findings 2024

双重二值化：1 + 1

■ DB-LLM首次实现了在2-bit条件下将困惑度降低到4.84

#Bits	Method	LLaMA-1-7B		LLaMA-1-13B		LLaMA-1-30B		LLaMA-1-65B	
		WikiText2	C4	WikiText2	C4	WikiText2	C4	WikiText2	C4
W16A16	-	5.68	7.08	5.09	6.61	4.10	5.98	3.53	5.62
W2A16†	RTN	188.32	151.43	101.87	76.00	19.20	30.07	9.39	11.34
W3A16	RTN	25.73	28.26	11.39	13.22	14.95	28.66	10.68	12.79
W2A16†	AWQ	2.5e5	2.8e5	2.7e5	2.2e5	2.3e5	2.3e5	7.4e4	7.4e4
W3A16	AWQ	11.88	13.26	7.45	9.13	10.07	12.67	5.21	7.11
W2A16†	GPTQ	22.10	17.71	10.06	11.70	8.54	9.92	8.31	10.07
W2A16†	OmniQuant	8.91	11.79	7.35	9.75	6.60	8.66	5.65	7.60
W2A16†	PB-LLM	20.61	47.09	10.73	25.40	9.65	16.28	6.50	11.13
W2A16†	DB-LLM	**7.59**	**9.74**	**6.35**	**8.42**	**5.52**	**7.46**	**4.84**	**6.83**

Table 1: **Performance comparisons of different methods for weight-only quantization on LLaMA-1** for language generation tasks. † represents the group size is 64.

- 在语言生成任务中优于其他量化方法
- 在超低比特量化（2-bit）条件下保持了较高的性能
- 在较大模型上的表现依然稳定

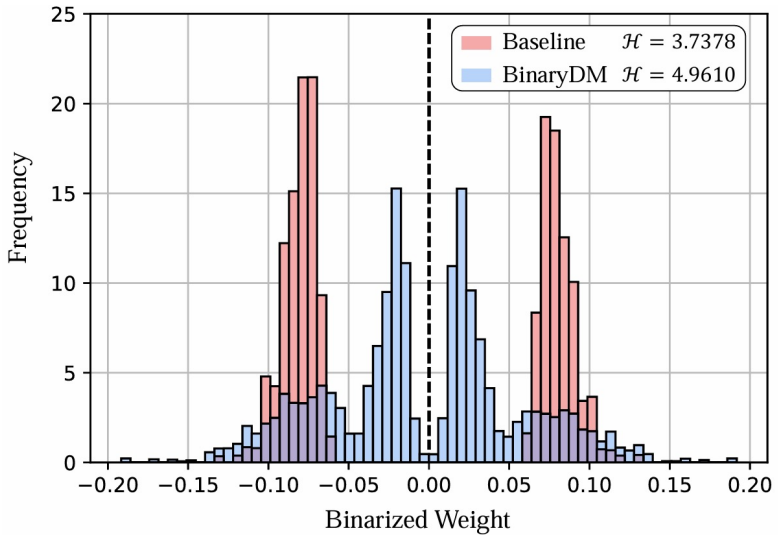
Model	#Bits	Method	Accuracy (%) ↑					
			PIQA	ARC-e	ARC-c	HellaSwag	Winogrande	Avg.
LLaMA-1-7B	W16A16	-	77.37	52.53	41.38	72.99	66.85	62.22
	W2A16	GPTQ	59.36	32.11	25.09	35.14	49.01	40.14
	W2A16	AWQ	50.05	25.76	29.44	25.93	49.96	36.23
	W2A16	OmniQuant	68.66	44.49	29.69	54.32	55.56	50.54
	W2A16	PB-LLM	55.39	34.22	24.23	31.99	52.88	39.74
	W2A16	DB-LLM	**72.14**	**44.70**	**33.62**	**60.71**	**61.01**	**54.44**

- 在零样本任务中的表现显著优于其他2-bit量化方法
- 在所有任务上的平均准确率最高

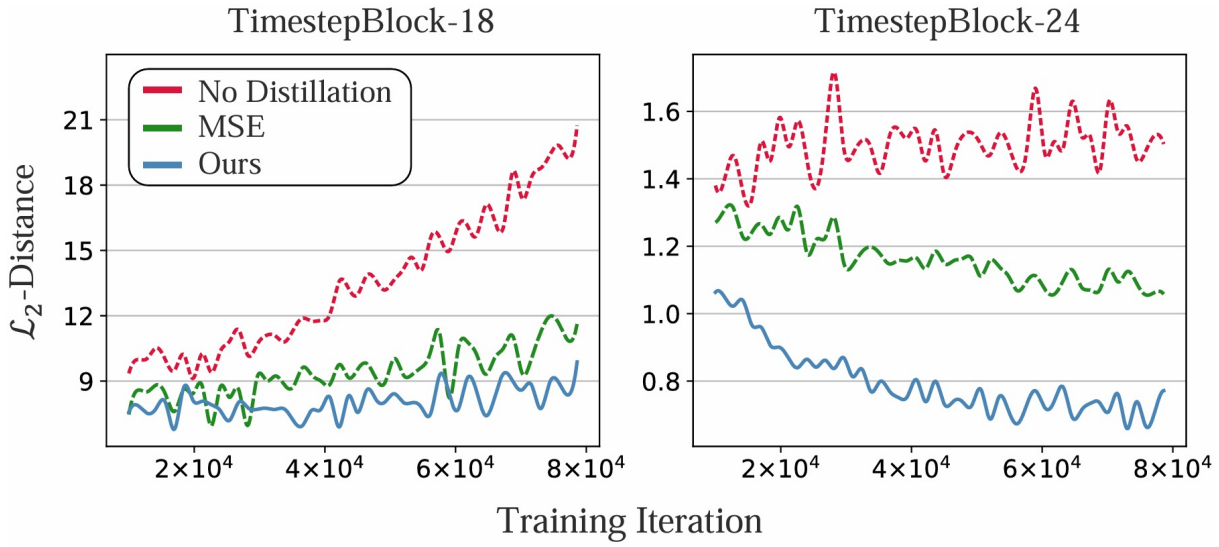
Chen et al., DB-LLM: Accurate Dual-Binarization for Efficient LLMs, ACL Findings 2024

Diffusion权重二值化

■ 问题：二值权重高度离散的表示方式会导致严重的精度下降



二值高度离散权重严重信息损失，阻碍生成表征



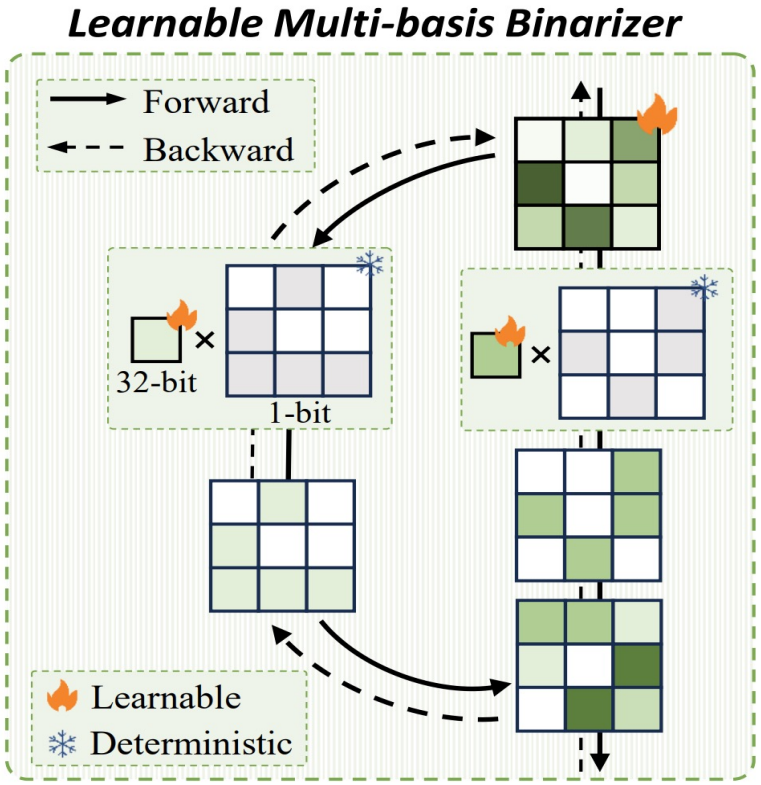
二值扩散模型训练优化方向难适应
训练损失函数难收敛

需要更准确的二值表征与优化方法

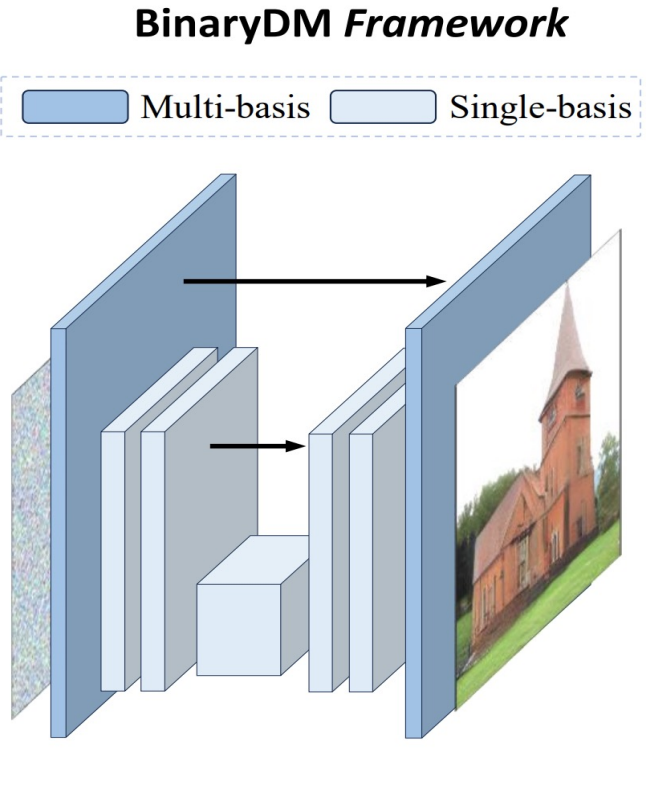
Zheng et al., Towards Accurate Binarization of Diffusion Model, NeurIPS 2024(submitted)

Diffusion 权重二值化

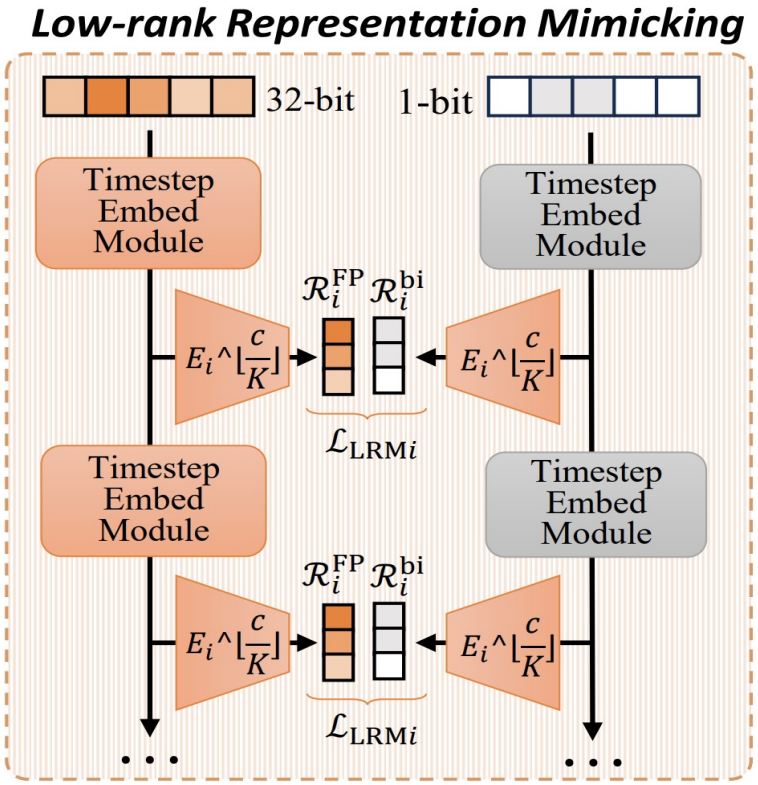
■ 思路：提出了一种针对扩散模型的量化感知训练方法BianryDM



可学习多基二值量化器
恢复二值生成表示



仅两端采用多基
小代价显著提升表达能力



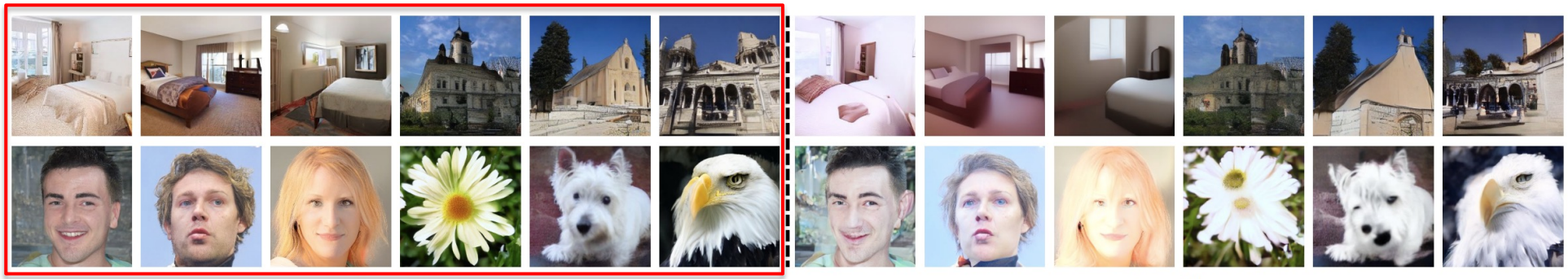
低秩表示模仿蒸馏
辅助优化二值扩散模型

解决了Diffusion在极限压缩场景下的表征能力损失问题

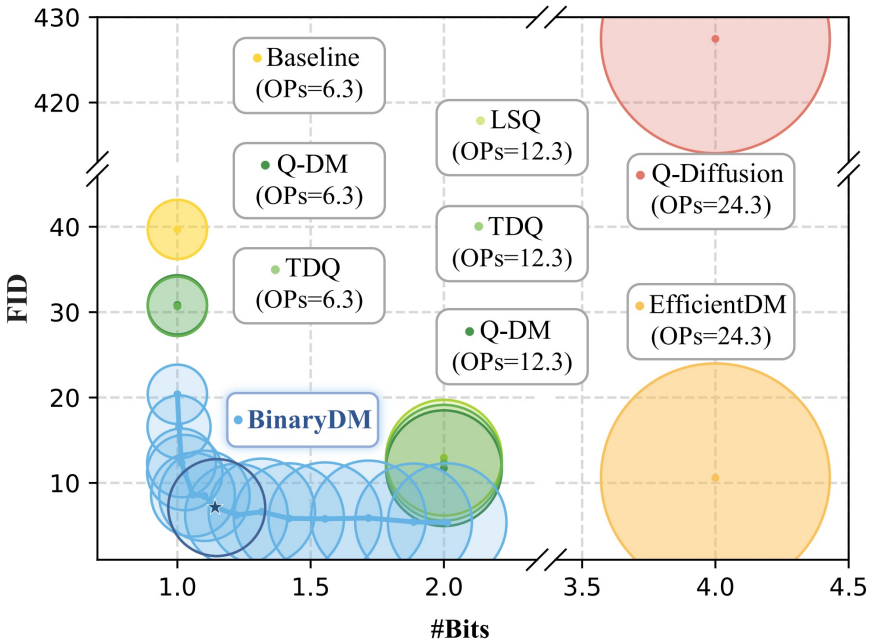
Zheng et al., Towards Accurate Binarization of Diffusion Model, NeurIPS 2024(submitted)

Diffusion 权重二值化

■ 首次实现二值Diffusion模型可用



BinaryDM　　Baseline



Model	Dataset	Method	#Bits	Size(MB)	FID↓	sFID↓	Precision↑	Recall↑
LDM-4	LSUN-Bedrooms 256 × 256	FP	32/32	1045.4	3.09	7.08	65.82	45.36
		LSQ	2/32	69.3	7.49	12.79	64.02	**37.60**
		Baseline	1/32	36.8	26.43	27.65	35.43	28.00
		BinaryDM	1/32	40.8	**6.42**	**9.63**	**66.92**	34.98
		Q-Diffusion	2/8	69.3	62.01	33.56	16.48	14.12
		LSQ	2/8	69.3	**6.08**	11.66	62.55	38.92
		Baseline	1/8	36.8	21.18	25.92	39.04	29.46
		BinaryDM	1/8	40.8	6.60	**11.63**	**67.23**	**33.48**
		EfficientDM	4/4	131.0	10.60	-	-	-
		Q-Diffusion	4/4	131.0	427.46	277.22	0.00	0.00
		LSQ	2/4	69.3	12.95	12.79	55.97	**34.30**
		Baseline	1/4	36.8	39.69	35.31	26.94	22.62
		BinaryDM	1/4	40.8	**7.11**	**11.26**	**62.21**	33.28

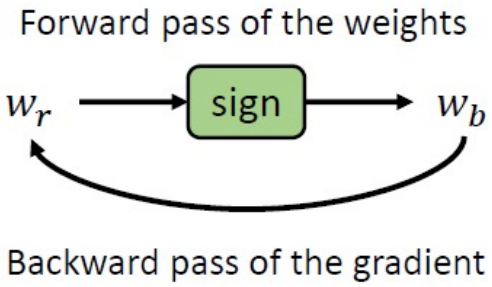
W1.1A4的BinaryDM达到了低至7.11的FID
实现了9.3×OPs和24.8×模型尺寸的减少

Zheng et al., Towards Accurate Binarization of Diffusion Model, NeurIPS 2024(submitted)

CUDA架构二值计算核

■ 问题：二值量化在GPU处理器上尚无底层系统支持

伪二值量化无法实际加速或节省显存



$$w_r = \alpha_w \mathbf{sign}(w_b)$$

$$w_r = \alpha_w \mathbf{bool}(w_b)$$

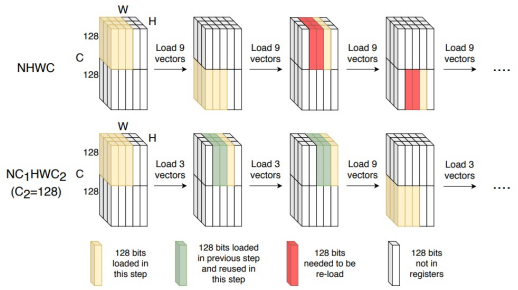
二值量化算法目前均使用伪量化形式模拟运算

$$\text{bi-linear}(\mathbf{X}) = \alpha_{\mathbf{w}}\alpha_{\mathbf{x}} \cdot (\text{sign}(\mathbf{X}) \otimes \text{sign}(\mathbf{W} - \mu(\mathbf{W})))$$

数值为1或-1，精度仍为FP16

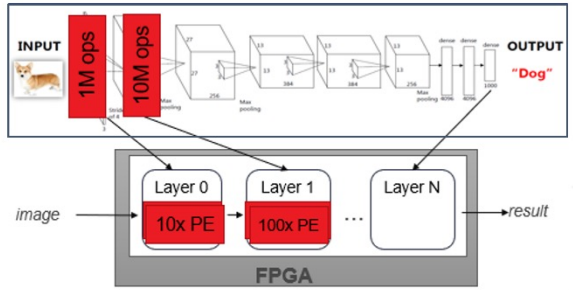
仅验证算法精度，而无法实际部署到GPU上

当前二值系统实现聚焦于移动端设备



dabnn仅支持二值卷积网络在ARM架构端上推理

FINN框架支持FPGA上的部分二值模型推理



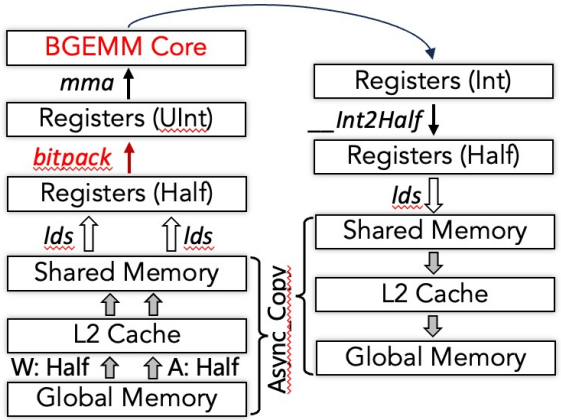
端侧设备存储和算力有限
无法支持更大模型推理

项目地址：https://github.com/yifu-ding/BGEMM-CUDA

CUDA架构二值计算核

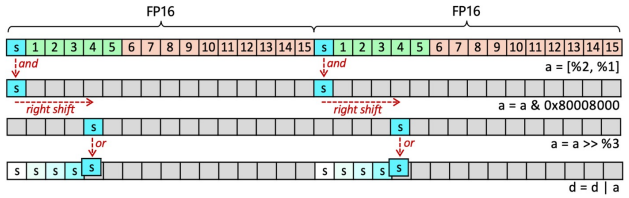
■ 思路：提出二值矩阵乘计算核，提高二值训练推理效率

二值加速器数据流

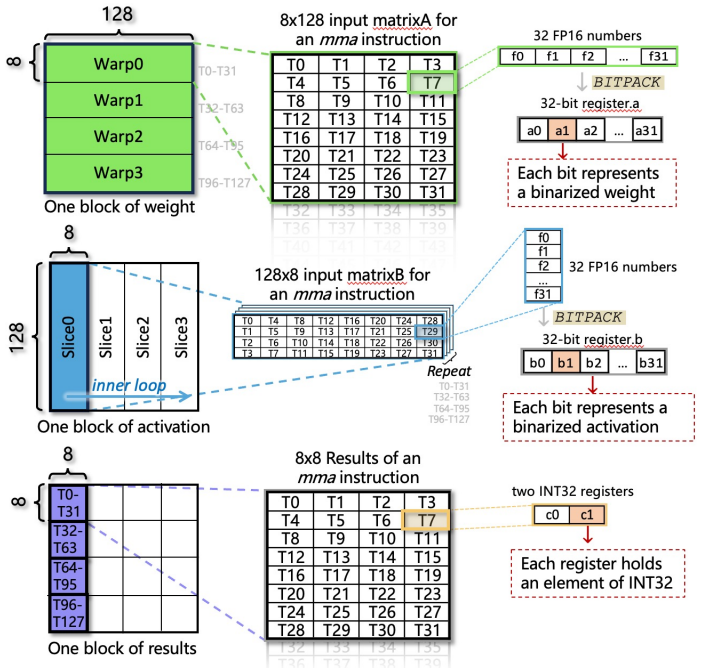


设置共享内存**缓冲区**，同时进行
“读”当前数据”和“写”下一轮数据

单指令**双路并行**二值量化单元

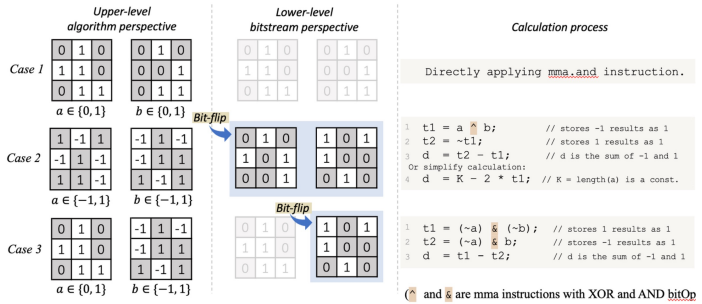


高效SIMT脉动数据流



数据**分片**加载，防止冗余操作
分块计算完毕后释放权重，**避免重复**量化
单指令多线程**高效并行**

支持多类二值化函数



实现**sign、bool**等二值量化形式
以支持各种上层算法设计

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases} \qquad g_X = \frac{d[X]}{dX} g_O[W^\top]$$

$$\text{bool}(x) = \begin{cases} 1, & x \geq 0.5 \\ 0, & x < 0.5 \end{cases} \qquad g_X = \frac{d[X]}{dX} g_O \frac{[W^\top] + 1}{2}$$

实现带反向传播的PyTorch接口，
支持训练

项目地址：https://github.com/yifu-ding/BGEMM-CUDA

CUDA架构二值计算核

■ 矩阵乘计算效率远超cuBLAS加速库中的FP16乘法实现3-5倍

M	N (batch size)	K	SplitK	Iteration	Time/ms			Performance/TFLOPs		
					cuBLAS	fp6_llm	BGEMM	cuBLAS	fp6_llm	BGEMM
128	32	128	1	10000	0.003	not supported	0.003	0.32	not supported	0.34
256	32	256	1	10000	0.009	0.008	0.004	0.45	0.5	1.19
1024	32	1024	1	10000	0.009	0.023	0.006	7.83	2.98	10.69
2048	32	2048	1	10000	0.023	0.039	0.009	11.86	6.81	30.03
13824	128	5120	1	10000	0.25	0.356	0.073	72.61	50.85	248.71
5120	128	13824	1	10000	0.237	0.479	0.087	76.38	37.83	208.87
22016	128	8192	1	10000	0.743	illegal memory	0.162	62.13	illegal memory	284.6
8192	128	22016	1	10000	0.732	illegal memory	0.148	63.05	illegal memory	312.75

二值计算核的计算效率约为同参数量的FP16矩阵乘计算核的3~5倍
展现出二值加速器的实际应用潜力

项目地址：https://github.com/yifu-ding/BGEMM-CUDA

目录

低比特量化

$$\mathbb{R} \longrightarrow \mathbb{I} \longrightarrow \mathbb{B} \in \{-1, +1\}$$

32-bit　　8-bit　　1-bit

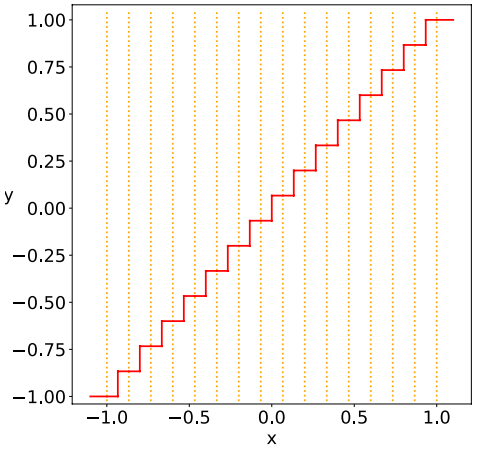
一、线性量化（2-8bit）

量化函数：

$$x_{int} = round(\frac{x}{\Delta}) + z$$

$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

反量化函数：$x_{float} = (x_Q - z)\Delta$



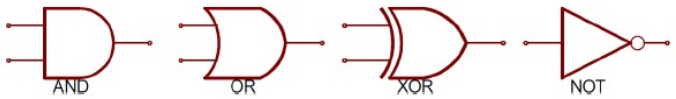
线性量化

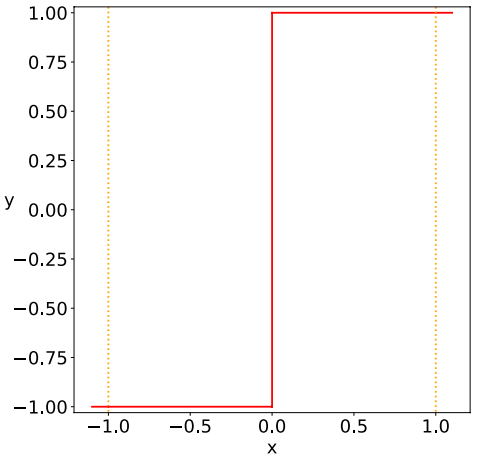
二、二值量化（1-bit）

量化函数：

$$Q_B(x) = \text{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

按位运算：

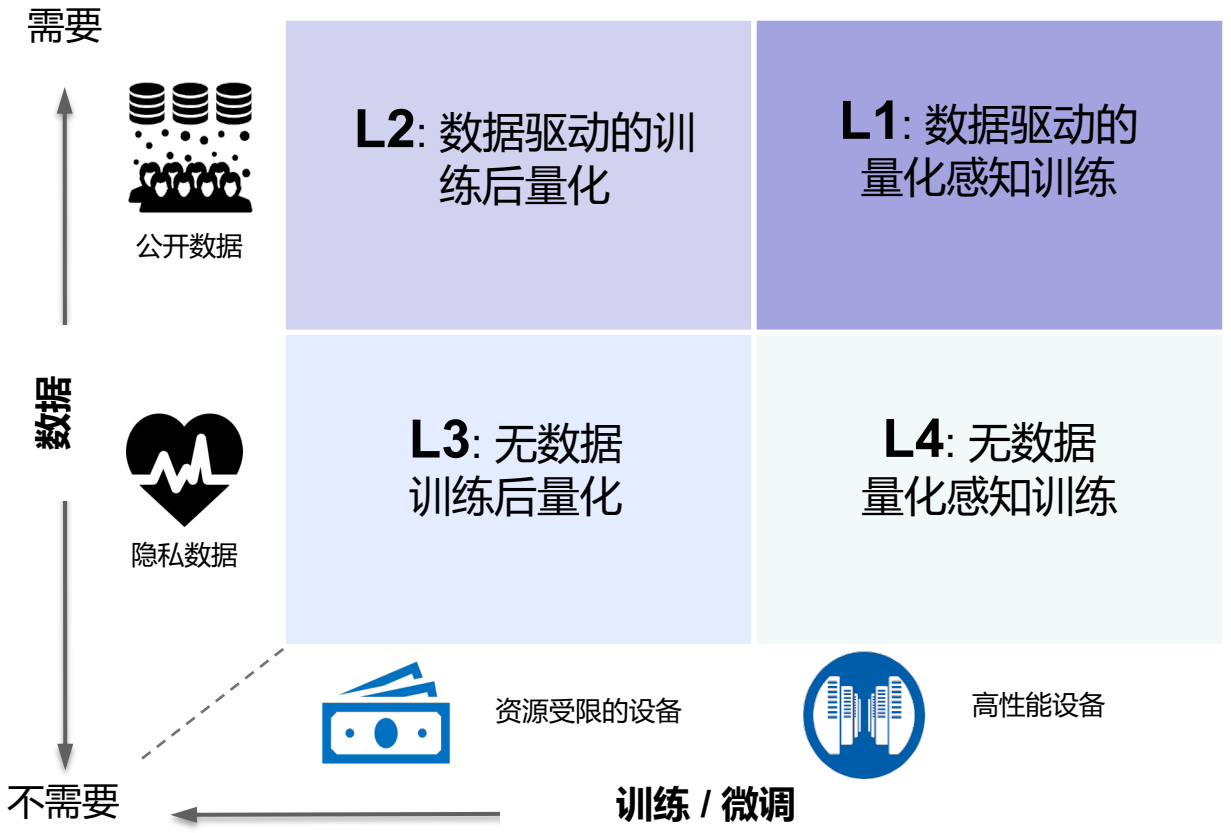




二值量化

低比特量化

- **挑战：无数据、低算力等限制场景下的模型量化生产和优化**



Post Training Quantization (PTQ)

Quantization Aware Training (QAT)

总结

多样场景

视觉内容生成

文本内容生成

Certainly! Hello, I'm ChatGPT, a state-of-the-art language model developed by OpenAI. I'm designed to assist you with a wide range of topics and provide helpful information to the best of my abilities. Whether you have questions, need creative ideas, or simply want to engage in a conversation, I'm here to help. Please keep in mind that while I strive to provide accurate and up-to-date information, my responses are based on my training data up until September 2021, so some information may be outdated. How can I assist you today?

需求约束

基础结构（Attention/SSM/...）
低/无数据　有限算力

训练加速　推理加速

复杂硬件

高性能集群

单机设备

边缘设备

展望

现有压缩研究	大模型特性	大模型压缩研究
理论体系完备	表征机理不完善	大模型压缩理论体系
硬件支持完善	模型结构异构	硬件友好的模型算子设计
无需解码策略	解码速度瓶颈	大模型高效解码策略