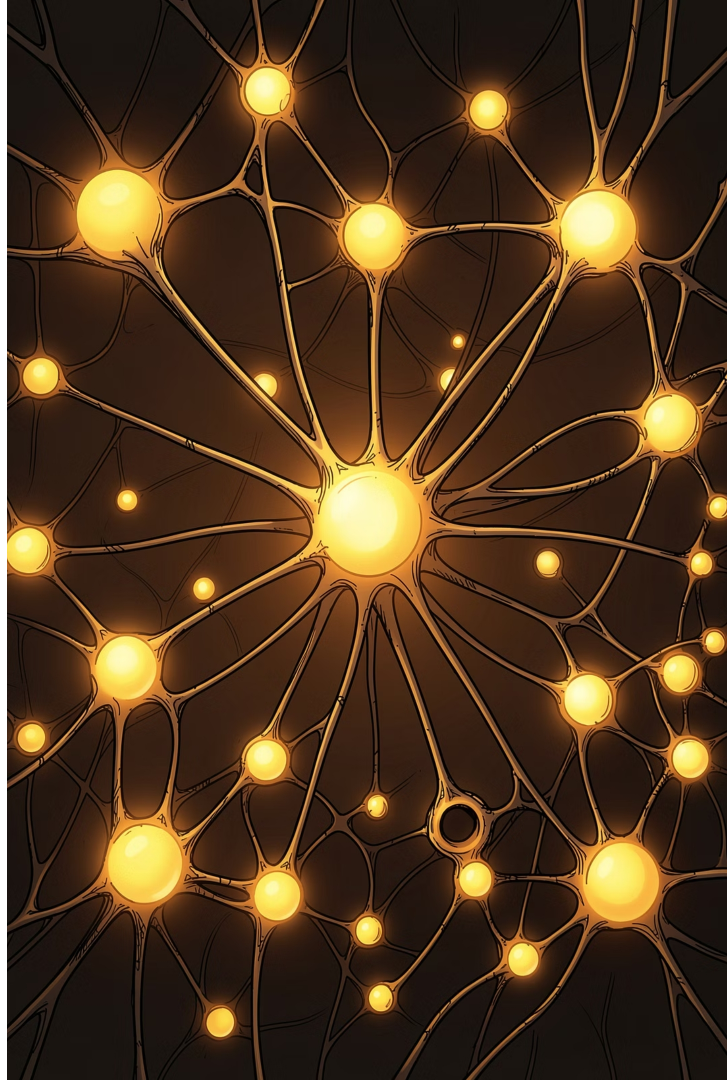


D²Quant: 面向 LLM 的低比特权重量化方法

作者: Xianglong Yan, Chengzhu Bao, Zhiteng Li, Tianao Zhang, Shaoqiu Zhang, Ruobing Xie, Xingwu Sun, Yulun Zhang

核心贡献: 提出 **Dual-Scale Quantizer + Deviation-Aware Correction**, 从权重和激活两个层面协同提升 sub-4-bit weight-only PTQ 精度。



现有问题与论文动机

前人方法的局限

- GPTQ / GPTAQ 在 4-bit 表现较好，但低于 4-bit 后精度明显下降
- Down-projection 矩阵对量化极为敏感，通常需要更高 bit-width 保护
- Weight-only PTQ 不量化激活，但权重量化仍会导致激活分布漂移

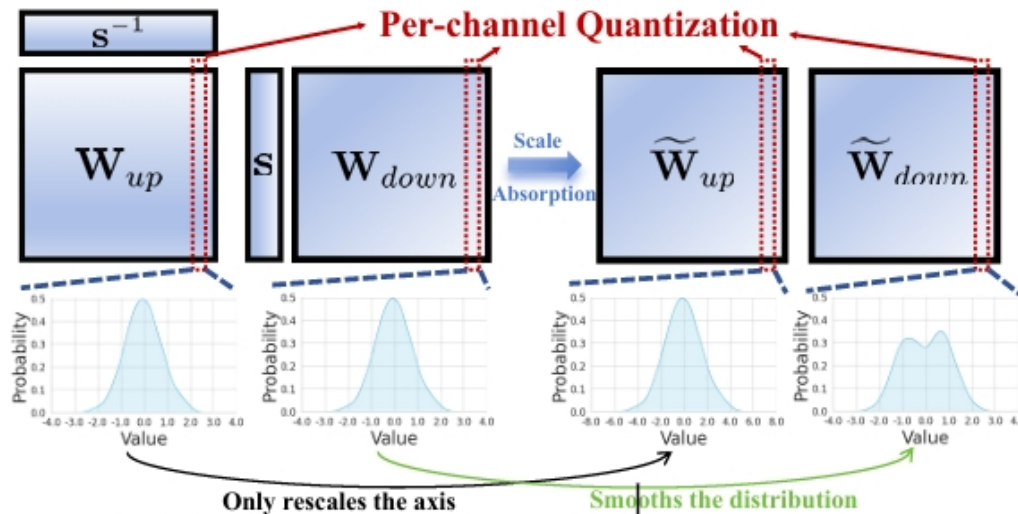
本文的创新应对

- 针对 2-bit / sub-4-bit 场景，重新系统分析误差来源
- 提出 **DSQ**，在不增加 bit budget 的前提下提升 down-projection 量化精度
- 提出 **DAC**，在 post-attention LayerNorm 中校正均值偏移

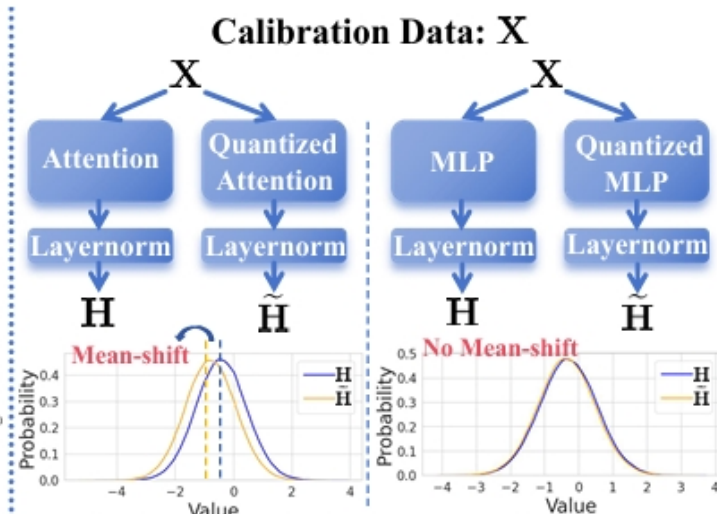


动机：让低比特权重量化既保持压缩优势，又尽量减少推理性能损失。

现有问题与论文动机



(a) An Equivalent Transformation Between Up and Down Projections



(b) Activation Deviation After Module Quantization

Figure 2. (a) Equivalent transformation between up and down projections in per-channel quantization: smoothing can be applied to down projection, while up projection only introduces channel-wise scaling. (b) Activation deviation at the subsequent LayerNorm after quantizing attention and MLP: quantizing attention causes a notable mean shift, whereas MLP quantization introduces no significant shift.

D²Quant 整体框架

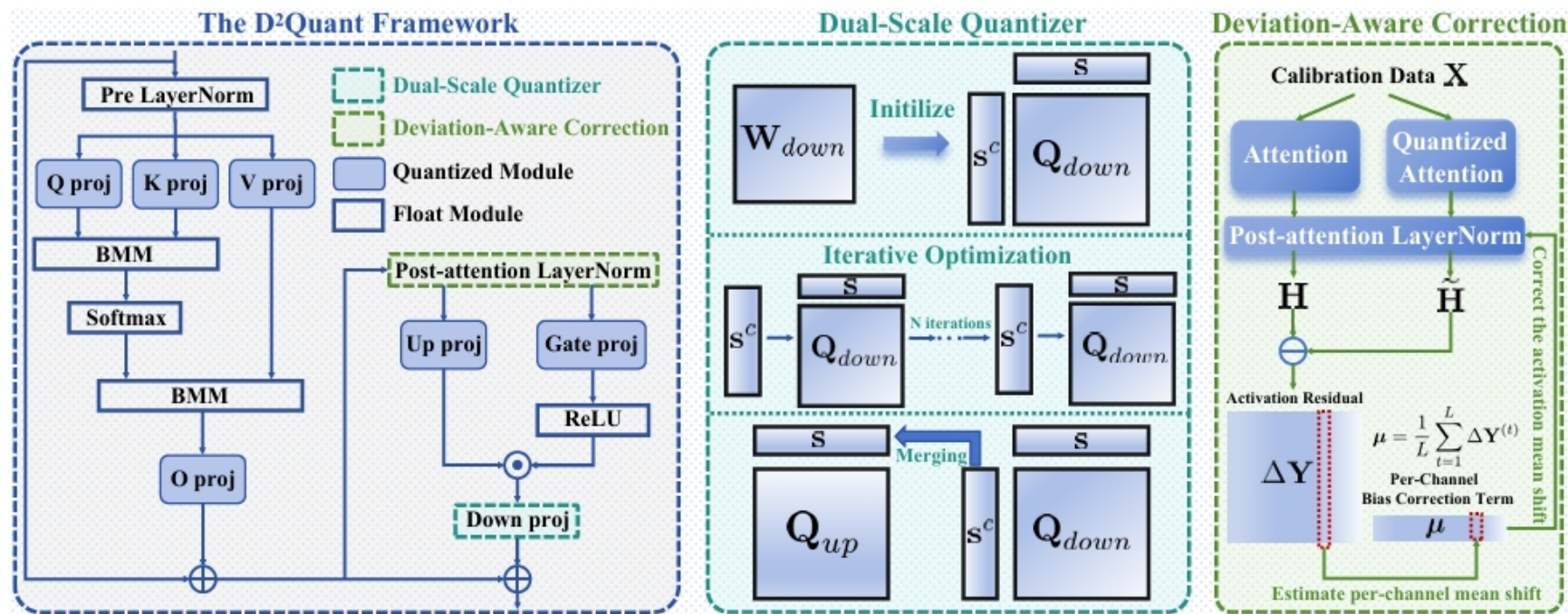
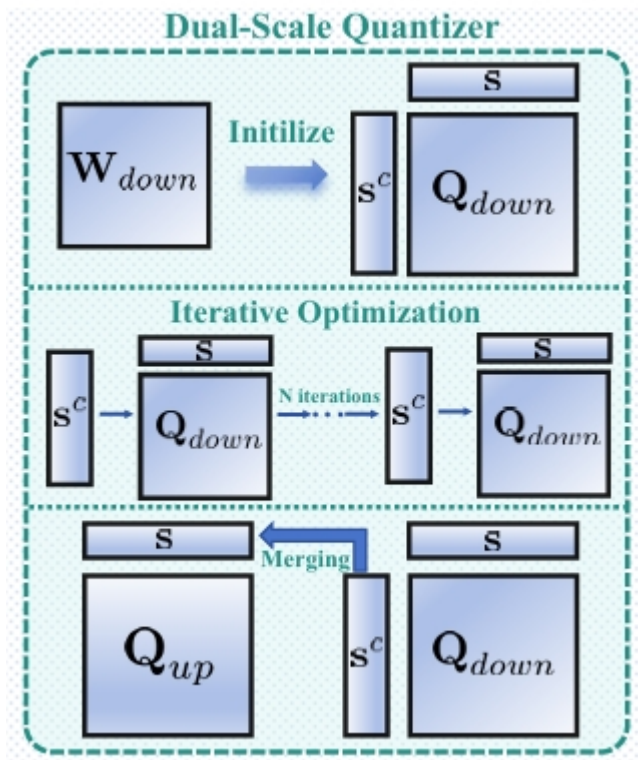


Figure 3. Overview of **D²Quant**. The left panel shows the D²Quant framework, which improves weight-only PTQ at both weight and activation levels. The middle shows the *Dual-Scale Quantizer*, which introduces an additional scale to refine the down-projection. The right depicts the *Deviation-Aware Correction*, which mitigates mean shift at post-attention LayerNorm via bias alignment.

核心方法一：Dual-Scale Quantizer (DSQ)



问题与观察

MLP 中 down-projection 矩阵是低比特量化瓶颈；up/down projection 之间可插入等价缩放变换。

DSQ 三步操作

1. 在 down-projection 量化中引入额外 **column-wise scale**
2. 通过迭代优化同时寻找量化权重与缩放因子
3. 将额外 scale 吸收到 up-projection 中，保持结构等价

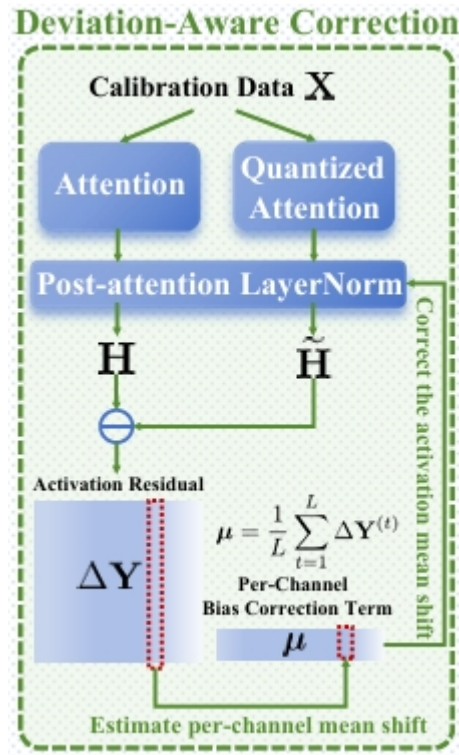
核心方法二：Deviation-Aware Correction (DAC)

问题发现

权重量化虽不直接量化激活，但会导致后续激活分布偏移。论文发现 **Attention 量化后，post-attention LayerNorm 出现明显 mean shift**，而 MLP 量化后的偏移相对不显著。

DAC 校正流程

1. 用校准数据估计 full-precision 与 quantized 输出的均值偏移 μ
2. 将 μ 作为 bias correction 加入 post-attention LayerNorm



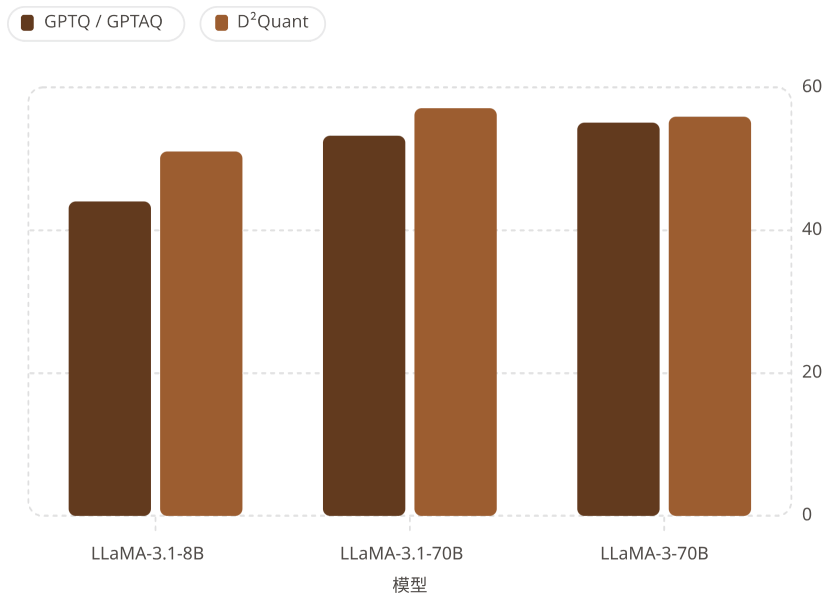
关键结果一：Qwen-3 系列 2-bit 量化

Table 1. Performance of 2-bit weight quantization on the Qwen-3 series. We report perplexity ($\downarrow$) on WikiText2 and C4, accuracy ($\uparrow$) on MMLU, and on seven commonsense tasks with their average (Avg.). $\dagger$ denotes methods with QuaRot rotation. Best results are in **bold**.

Model	Method	Wiki2($\downarrow$)	C4($\downarrow$)	MMLU($\uparrow$)	PiQA	Hella.	Arc-E	Arc-C	Wino.	RTE	OBQA	Avg.($\uparrow$)
Qwen3 8B	FP16	9.72	15.43	72.96	77.64	74.90	80.81	57.00	68.03	77.98	41.80	68.31
	GPTQ	23.28	55.55	37.05	65.23	47.27	48.11	31.14	55.56	57.04	30.40	47.82
	GPTQ †	15.27	28.60	46.53	66.92	53.03	56.52	35.92	61.88	71.48	32.60	54.05
	GPTAQ	20.24	52.02	30.37	61.81	42.05	45.88	28.67	58.33	60.29	29.60	46.66
	GPTAQ †	15.61	29.91	43.49	66.81	49.42	54.80	33.02	61.64	64.26	31.40	51.62
	BoA	17.67	40.37	40.37	65.45	46.93	55.18	32.34	59.75	54.51	31.2	49.34
	D ² Quant	14.10	25.96	49.94	70.51	55.76	66.84	39.25	61.96	71.84	34.40	57.22
Qwen3 14B	FP16	8.64	13.82	77.11	79.76	78.89	83.04	60.24	73.01	77.62	46.80	71.34
	GPTQ	12.71	28.68	53.20	71.93	61.87	62.84	39.08	64.72	49.46	36.40	55.19
	GPTQ †	12.92	38.23	46.90	71.98	54.43	67.30	40.78	66.30	66.43	37.40	57.80
	GPTAQ	12.91	28.19	51.33	71.06	60.46	66.58	41.47	66.77	61.73	39.00	58.15
	GPTAQ †	12.48	23.93	56.12	71.55	58.85	64.73	40.36	65.90	81.59	35.00	59.71
	BoA	14.10	29.33	44.95	70.24	54.25	58.04	35.49	65.51	74.01	32.80	55.76
	D ² Quant	11.88	22.40	58.36	72.31	60.98	71.51	44.54	66.85	74.73	39.20	61.45
Qwen3 32B	FP16	7.61	12.45	80.76	82.05	82.62	83.21	61.26	73.01	76.17	46.00	72.05
	GPTQ	11.13	25.27	63.04	73.78	68.76	64.06	44.71	68.75	74.01	40.20	62.04
	GPTQ †	10.61	19.38	63.03	72.52	68.80	67.34	48.12	67.88	67.87	41.60	62.02
	GPTAQ	10.78	24.12	62.54	73.94	68.33	66.84	44.62	69.06	67.51	38.20	61.21
	GPTAQ †	10.44	20.03	62.61	74.32	68.50	67.85	46.33	68.51	69.31	39.80	62.09
	BoA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	D ² Quant	9.71	17.10	64.31	76.12	70.47	75.51	51.11	70.09	68.23	40.40	64.56

关键结果二：LLaMA 系列 2-bit 量化

核心发现



- **LLaMA-3.1-8B**: 平均准确率提升 **+6.99**，提升幅度最为显著
- **LLaMA-3.1-70B**: C4 PPL 从 248.64 降至 **17.74**，语言建模能力大幅恢复
- **LLaMA-3-70B**: 平均准确率 55.08 → 55.91，稳定提升

□ D²Quant 不只适用于 Qwen，也能有效迁移到 LLaMA 系列，展现良好泛化性。

消融与对比实验结论

Table 3. Effect of DSQ and DAC Components. Component-wise breakdown analysis of DSQ and DAC on Qwen-3-8B (2-bit).

Model	Method	Wiki2(↓)	C4(↓)	MMLU(↑)	Acc(↑)
Q3-8B	Baseline	14.72	27.47	45.49	53.94
	+DSQ	14.41	26.95	48.97	57.21
	+DAC	14.63	27.00	45.82	54.63
	+DSQ+DAC	14.10	25.96	49.94	57.22

- **DSQ** 是主要性能提升来源 (+3.48) ，DAC 提供额外补偿 (+0.33)
- 两者结合达到最佳效果 **49.94**，验证协同有效性

局限性与未来方向

当前局限

- DSQ 目前主要针对 down-projection，尚未系统扩展至 Q/K/V、up/gate 等矩阵
- DAC 主要校正 mean shift，对其他类型激活分布漂移覆盖不足
- 方法依赖校准数据，校准集过小可能导致 correction 不稳定

未来研究方向

扩展 DSQ 适用范围

将 DSQ 推广至更多投影矩阵，解决多分支共享 scale 的约束问题

更通用的激活校正

识别更多量化诱发的激活偏移模式，设计更通用的 correction 机制