

R2Q: Towards Robust 2-Bit Large Language Models via Residual Refinement Quantization

Jiayi Chen¹ Jieqi Shi^{1,*} Jing Huo¹ Chen Wu²

¹Nanjing University ²Microsoft AI

Background

- 大语言模型（LLMs）的快速发展带来了巨大的计算和内存需求，推动了低比特量化的应用。
- 但将量化扩展到2比特仍然具有挑战性，因为会导致严重的精度下降。

Problem

- 残差精细量化 (R2Q) —— 一种新颖的2比特量化框架，将过程分解为两个连续的1比特量化，形成自适应量化格网。

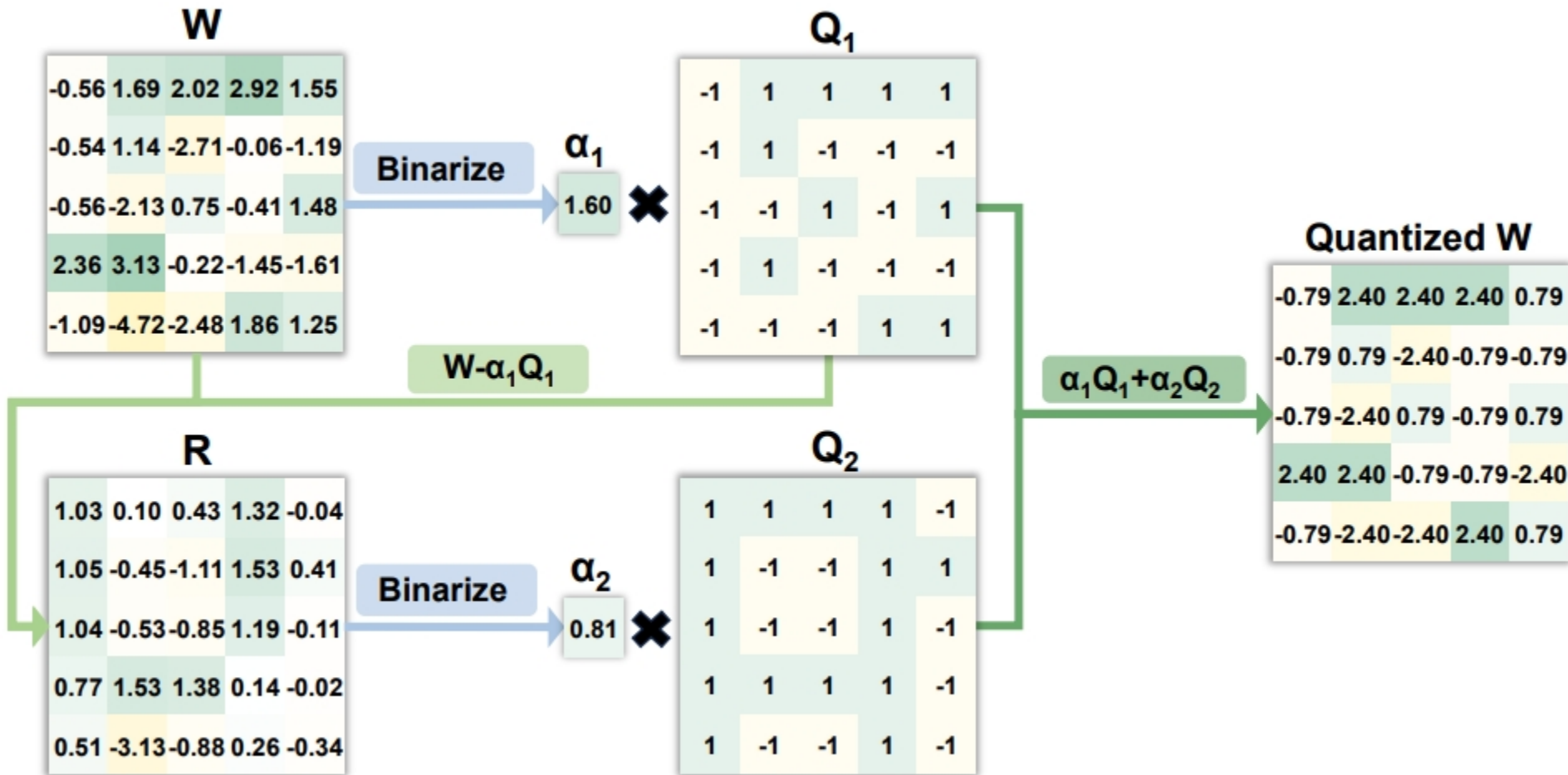
Existing solutions

- Quip: 2-bit quantization of large language models with guarantees

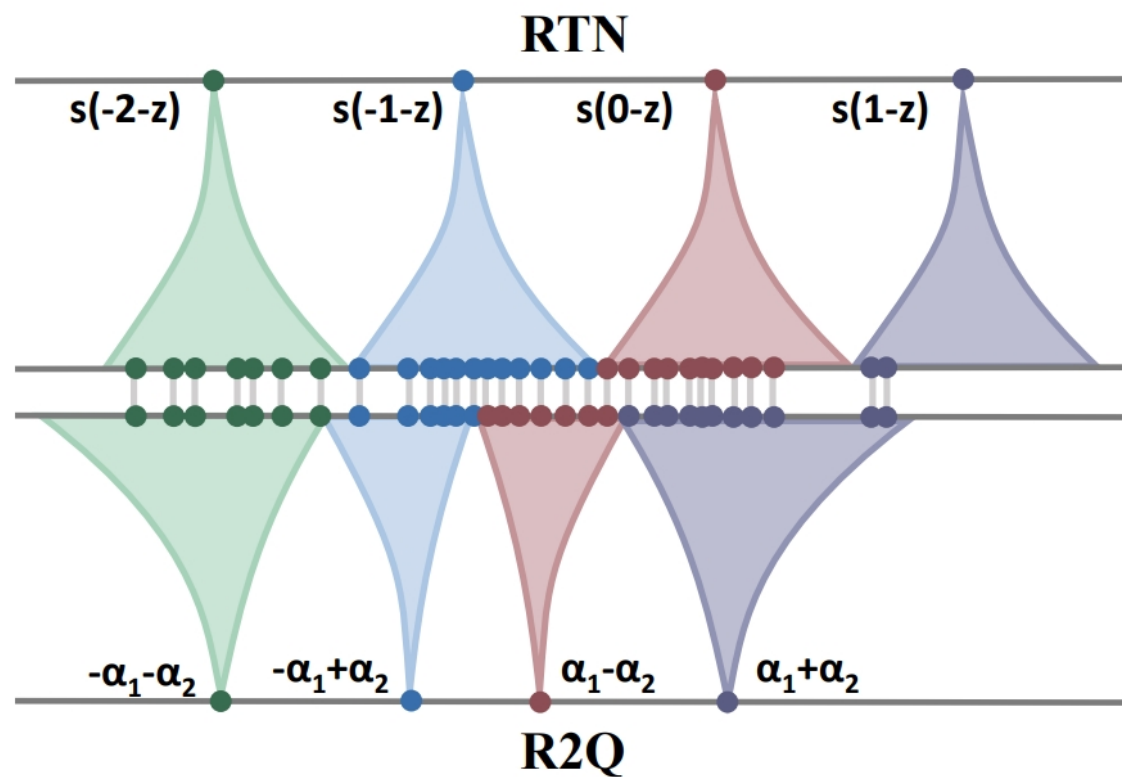
QUIP 将量化误差建模为一个二阶优化问题，通过最小化 $(W - Q)^T H(W - Q)$ ，利用 Hessian 捕捉参数的重要性，从而在极低比特下保持模型性能。

$$L(W + \Delta W) \approx L(W) + \frac{1}{2} \Delta W^T H \Delta W$$

Basic idea



Method

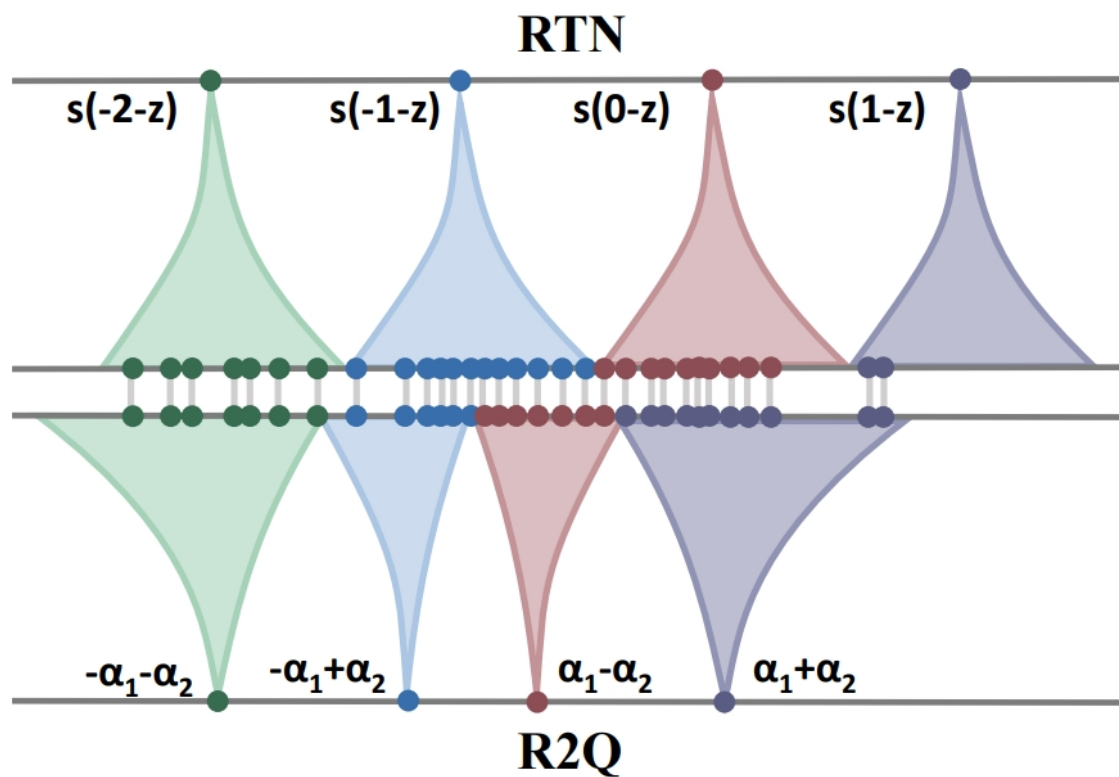


$$s^{(i)} = \frac{r_{\max}^{(i)} - r_{\min}^{(i)}}{q_{\max} - q_{\min}},$$

$$q_{\min} = \frac{r_{\min}}{s} + z \quad z^{(i)} = \left\lfloor \frac{q_{\min} - r_{\min}^{(i)}}{s^{(i)}} \right\rfloor,$$

$$q^{(i)} = \mathcal{Q}_{s^{(i)}, z^{(i)}}(w^{(i)}) = \text{clip} \left(\left\lfloor \frac{w^{(i)}}{s^{(i)}} \right\rfloor + z^{(i)}, q_{\min}, q_{\max} \right),$$

Method



$$\|\mathbf{w} - \alpha \mathbf{q}\|_F^2 = \|\mathbf{w}\|_F^2 - 2\alpha \langle \mathbf{w}, \mathbf{q} \rangle_F + \alpha^2 \|\mathbf{q}\|_F^2 ,$$

$$\frac{\partial}{\partial \alpha} (\|\mathbf{w}\|_F^2 - 2\alpha \langle \mathbf{w}, \mathbf{q}^* \rangle_F + G\alpha^2) = -2\langle \mathbf{w}, \mathbf{q}^* \rangle_F + 2G\alpha = 0 .$$

$$\alpha^* = \frac{\langle \mathbf{w}, \mathbf{q}^* \rangle_F}{G} = \frac{\sum_i w_i \text{sign}(w_i)}{G} = \frac{\sum_i |w_i|}{G} = \frac{\|\mathbf{w}\|_{\ell_1}}{G} ,$$

$$\alpha_1 = \frac{\|W\|_1}{G}, \quad \alpha_2 = \frac{\|R\|_1}{G}$$

$$W \approx \alpha_1 q_1 + \alpha_2 q_2$$

Experiments

Model	Bit-Width	Method	Group-Size	CR	ARC-c	ARC-e↑	BoolQ ↑	Hella. ↑	PIQA↑	Wino.↑	MMLU↑	Wiki2↓
Llama-7B	bf16	\	\	100.00%	41.98	75.38	75.17	56.95	78.78	69.46	31.33	9.39
	2-bit	LLM-QAT	-1	12.3%	19.45	26.81	37.83	25.71	51.90	50.28	22.95	1776.98
		BitDistiller	-1	12.3%	23.38	25.88	53.88	25.59	51.46	50.43	24.61	15938.61
		R2Q (ours)	-1	12.3%	27.47	56.82	59.36	44.44	70.08	57.54	24.08	17.13
		LLM-QAT	128	15.1%	26.19	42.00	58.32	35.32	62.13	54.22	23.91	115.52
		BitDistiller	128	15.1%	28.67	57.91	63.18	41.76	69.10	60.22	24.26	33.51
		R2Q (ours)	256	15.0%	28.24	59.18	64.89	45.37	70.62	59.59	24.28	16.96
OPT-6.7B	bf16	\	\	100.00%	30.38	65.53	65.72	50.53	76.22	64.88	24.94	12.28
	2-bit	LLM-QAT	-1	12.5%	19.03	38.51	60.64	30.34	60.39	50.59	22.92	53.24
		BitDistiller	-1	12.5%	21.50	42.55	48.69	31.88	61.10	52.41	24.57	125.01
		R2Q (ours)	-1	12.5%	25.68	58.08	63.7	44.19	71.43	60.77	24.60	16.71
		LLM-QAT	128	13.5%	19.71	39.23	59.97	36.17	61.75	51.01	25.63	31.07
		BitDistiller	128	13.5%	26.37	53.70	62.35	41.22	68.66	56.99	25.40	22.49
		R2Q (ours)	256	13.4%	24.32	58.96	64.40	44.31	72.58	60.54	25.09	16.81

Experiments

Model	Method	Group-Size	CR	Δ ARC-c	Δ ARC-e	Δ BoolQ	Δ Hella.	Δ PIQA	Δ Wino.	Δ MMLU	Δ Wiki2
Llama-7B	LLM-QAT	g128→g-1	15.1%→12.3%	-6.74	-15.19	-20.49	-9.61	-10.23	-3.94	-0.96	1661.46
	BitDistiller	g128→g-1	15.1%→12.3%	-5.29	-32.03	-9.30	-16.17	-17.64	-9.79	0.35	15905.10
	R2Q (ours)	g256→g-1	15.0%→12.3%	-0.77	-2.36	-5.53	-0.93	-0.54	-2.05	-0.20	0.17
OPT-6.7B	LLM-QAT	g128→g-1	13.5%→12.5%	-0.68	-0.72	0.67	-5.83	-1.36	-0.42	-2.71	22.17
	BitDistiller	g128→g-1	13.5%→12.5%	-4.87	-11.15	-13.66	-9.34	-7.56	-4.58	-0.83	102.52
	R2Q (ours)	g256→g-1	13.4%→12.5%	1.36	-0.88	-0.70	-0.12	-1.15	0.23	-0.49	-0.10

双击此处添加标题

THANK

S