IMPERIAL

Advanced Research + Invention Agency ARIA DeepWok Lab

QERA
An Analytical Framework For Quantization Error Reconstruction

Cheng Zhang
Jeffrey T. H. Wong, Can Xiao
A. Constantinides, Yiren Zhao

Quantization + Low-rank approximation

Compression

- Given a trained layer

$$y = xW$$

- Quantize

$$\mathrm{W_q} = \mathrm{quantize}(W)$$

- Then approximate error with low-rank terms:

$$(A_k, B_k) = \mathrm{approx}\big(W - W_q\big)$$

Inference

$$y = xW_q + xA_kB_k,$$

W_q is high-rank low-precision,

A_k, B_k are low-rank high-precision.

Applications

- Quantized parameter-efficient fine-tuning (QPEFT)
- Post-training quantization (PTQ)

Existing Methods

Problem to solve

How to approximate $y = xW$ with $y = x(W_q + C_k)$?

- Ignore error: qLoRA
- Minimize weight error

$$\mathrm{argmin_{C_k}} \|W - W_q - C_k\|_F^2$$

e.g, LoftQ, based on SVD on weights $\mathrm{C_k = SVD_k}(W - W_q)$

- Minimize output error

$$\mathrm{argmin_{C_k}} \mathrm{E_{y\sim Y}} \| y - \tilde{y}\|_2^2 \Rightarrow \mathrm{argmin_{C_k}} E_{x\sim X} \big\| x\big(W - W_q - C_k\big)\big\|_2^2$$

e.g, LQER, LQ-LoRA, based on heuristics

QERA

QERA provides closed-form solutions to minimizing output error

QERA-exact (exact solution)

$$C_k = \left(R_{XX}^{\frac{1}{2}}\right)^{-1} SVD_k\big(R_{XX}^{\frac{1}{2}}\big(W - W_q\big)\big)$$

where $R_{XX} \coloneqq E_{x\sim X}[x^T x]$ is the auto-correlation matrix of input vector space

QERA-approx (approximate solution)

$$C_k = S^{-1} SVD_k\big(S\big(W - W_q\big)\big)$$

where $S = diag\left(\sqrt{E[x \odot x]}\right)$, $\odot$ denotes elementwise multiply.

Evaluation: Improved QPEFT & PTQ

QPEFT fine-tuning RoBERTa-base on GLUE

W-bits	Rank	Method	MNLI Acc	QNLI Acc	RTE Acc	SST Acc	MRPC Acc	CoLA Matt	QQP Acc	STSB P/S Corr	Avg.
16	-	Full FT	87.61	92.95	73.16	94.88	92.15	60.41	91.61	90.44/90.25	85.38
16	8	LoRA	87.85	92.84	69.55	94.46	89.99	57.52	89.83	89.92/89.83	84
4.25	8	QLoRA	87.21	92.32	63.9	94.08	88.24	56.08	90.55	89.59/89.56	82.75
		LoftQ (5-iter)	87.27	92.48	67.13	94.38	88.24	54.59	90.51	88.75/88.79	82.92
		QERA-approx	87.28	92.45	70.4	94.38	88.97	55.99	90.39	89.83/89.72	83.71
3.25	8	QLoRA	84.87	89.58	53.67	91.02	73.94	3.12	89.31	84.80/84.38	71.29
		LoftQ (5-iter)	85.24	89.65	58.24	92.05	75.82	11	88.93	85.55/85.27	73.31
		QERA-approx	85.58	90.74	58.48	92.59	82.19	32.98	89.41	87.43/87.08	77.43
2.5	64	QLoRA	77.87	85.26	54.15	90.02	71	0	87.93	74.72/75.31	67.62
		LoftQ (5-iter)	80.15	87.65	52.95	90.94	74.35	3.43	89.17	82.76/82.90	70.18
		QERA-exact	84.64	90.05	58.48	92.32	84.72	26.43	89.69	86.48/86.40	76.23

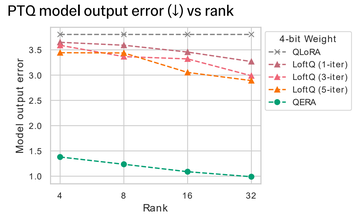
PTQ perplexity (↓) across various models

W-bits	Rank	Method	TinyLlama 1.1B	Gemma-2 2B	Phi-3.5 3.8B	LLaMA-2 7B	LLaMA-2 13B	LLaMA-3.1 8B	LLaMA-3.1 80B
-	-	BF16	13.98	13.08	11.5	8.71	7.68	7.55	3.06
4.25	-	HQQ	15.02	14.29	14.63	9.59	8.27	8.72	3.97
4.25	32	w-only	19.4	16.23	14.16	9.45	8.06	8.78	4.55
		ZeroQuant-V2	18.03	15.71	14.09	9.42	8.07	8.83	4.48
		LQER	16.23	14.55	12.88	9.22	7.96	8.45	4.1
		QERA-approx	15.66	14.6	12.81	9.17	7.95	8.45	4.1
		QERA-exact	16.16	14.12	12.3	9.12	7.93	8.33	3.82
3.25	64	w-only	32.82	41.13	47.78	13.32	10.24	18.96	16.46
		ZeroQuant-V2	27.8	33.56	42.64	13	10.03	19.29	10.12
		LQER	20.6	21.99	18.27	14	9.09	11.86	7.05
		QERA-approx	20.43	21.93	17.99	10.99	9.04	11.73	6.99
		QERA-exact	19.51	19.97	20.37	10.67	8.97	11.39	6.68

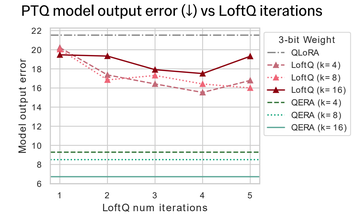
How does QERA work?

Layer weight error vs layer output error vs model output error?

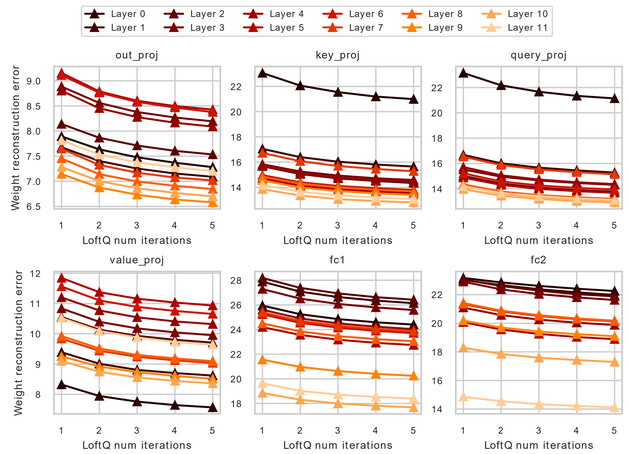
PTQ model output error (↓) vs rank



PTQ model output error (↓) vs LoftQ iterations

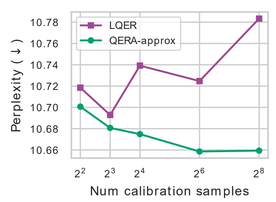


PTQ layer weight approximation error (↓) vs LoftQ iterations

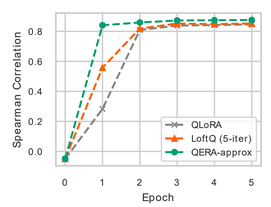


Better stability, higher efficiency

PTQ perplexity (↓) vs calibration set size



QPEFT fine-tuned spearman correlation (↑) vs epochs



Find more details in our paper and open-sourced codes (QR code on the top right of the poster)