UNIVERSITY AT ALBANY
State University of New York

MagR: Weight Magnitude Reduction for Enhancing Post-Training Quantization

Aozhong Zhang[1], Naigang Wang[2], Yanxia Deng[1], Xin Li[1], Zi Yang[1], Penghang Yin[1]

[1]University at Albany, SUNY [2]IBM T. J. Watson Research Center

UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

IBM

Introduction

Large Language Models (LLMs) perform well but are limited by their size and computational demands. Low-precision Post-Training Quantization (PTQ) helps reduce these demands, though it may slightly affect accuracy compared to Quantization Aware Training (QAT) at very low precision. Recent PTQ advancements use **linear transformations** to make pre-trained weights easier to quantize by reducing magnitudes and suppressing outliers. In a nutshell, given the features $\boldsymbol{X}$ and weights $\boldsymbol{W}$, a linear transformation $\boldsymbol{T}$ is applied to weights $\boldsymbol{W}$ to make $\boldsymbol{TW}$ more quantization-friendly than $\boldsymbol{W}$. i.e.

$$\boldsymbol{XW} = (\boldsymbol{XT}^{-\mathbf{1}})(\boldsymbol{TW}) \approx (\boldsymbol{XT}^{-\mathbf{1}})\boldsymbol{Q}(\boldsymbol{TW})$$

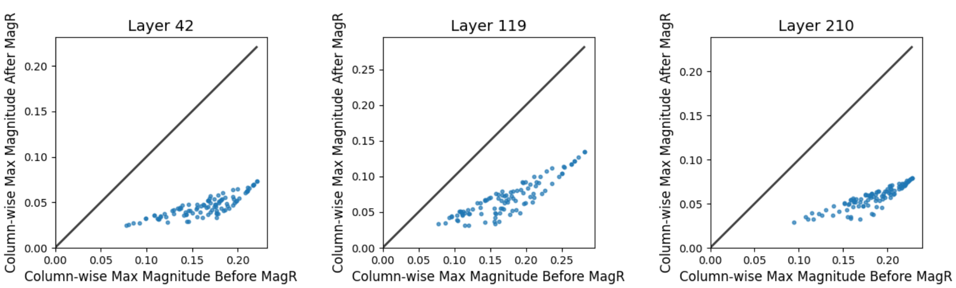
$\boldsymbol{Q}(\boldsymbol{TW})$ is the quantized weights. However, this PTQ method requires architecture changes and extra inference overhead. To overcome this, we propose MagR, a non-linear method with channel-wise l_∞-regularized least squares, reduces quantization scale without adding inference overhead.

$$\min_{\boldsymbol{W}\in\mathbb{R}^m} \frac{1}{2}\left\|\boldsymbol{XW} - \boldsymbol{X}\widehat{\boldsymbol{W}}\right\|^2 + \alpha\|W\|_\infty$$

where α serves as the regularization parameter, balancing fidelity and regularization.

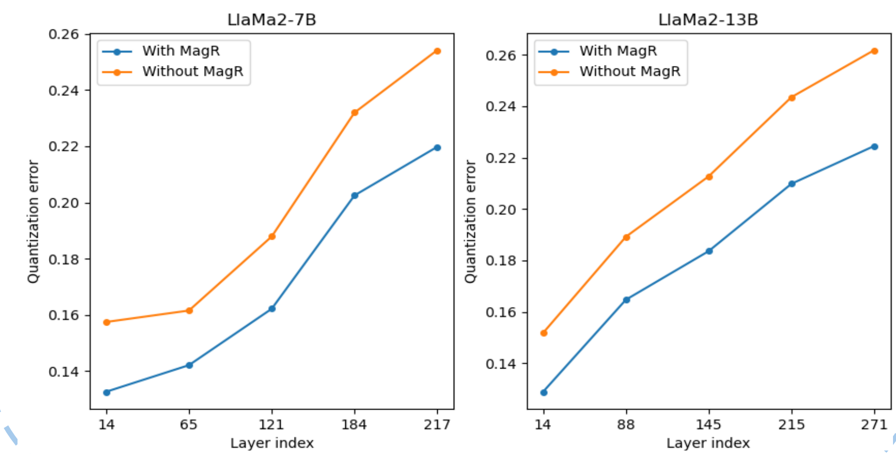
To address the l_∞-regularization problem, we develop a parallelizable proximal gradient descent algorithm that computes l_1-ball projections at each iteration. MagR achieves two key goals:

- It reduces the channel-wise maximum weight magnitude.



- It preserves model performance with minimal accuracy loss and reduces quantization error.

Model	Method	Wikitext2 (PPL↓)	C4 (PPL↓)
LLaMA2-7B	Original	5.47	6.97
	MagR	5.52	7.04
LLaMA2-13B	Original	4.88	6.46
	MagR	4.92	6.52
LLaMA2-70B	Original	3.31	5.52
	MagR	3.35	5.56



UNIVERSITY AT ALBANY
State University of New York

Method

Proximal Gradient Descent

l_∞-norm is convex but not differentiable. Instead of subgradient methods, we use proximal gradient algorithm to solve it, which matches the convergence rate of standard gradient descent. With the step size $\eta > 0$, proximal gradient descent takes the following iteration:

$$\boldsymbol{w}^{k+1} = \operatorname{prox}_{\eta\alpha\|\cdot\|_\infty}\left(\boldsymbol{w}^k - \eta\nabla_{\boldsymbol{w}}\frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}\widehat{\boldsymbol{w}}\|^2\right)$$

$$= \operatorname{prox}_{\eta\alpha\|\cdot\|_\infty}\left(\boldsymbol{w}^k - \eta\cdot\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{X}(\boldsymbol{w}^k - \widehat{\boldsymbol{w}})\right)$$

where $\operatorname{prox}_{t\|\cdot\|_\infty}$ with the scalar $t > 0$ is the (scaled) proximal operator of l_∞-norm function, defined as:

$$\operatorname{prox}_{t\|\cdot\|_\infty}(\boldsymbol{v}) := \arg\min_{\boldsymbol{x}\in\mathbb{R}^m}\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|^2 + t\|\boldsymbol{x}\|_\infty$$

Proximal Operator of l_∞-Norm

It turns out we can compute it by leveraging the celebrated Moreau decomposition

$$\operatorname{prox}_{t\|\cdot\|_\infty}(\boldsymbol{v}) := \boldsymbol{v} - t\cdot\operatorname{proj}_{\|\cdot\|_1\le 1}\left(\frac{\boldsymbol{v}}{\boldsymbol{t}}\right)$$

$$\operatorname{proj}_{\|\cdot\|_1\le 1}(\boldsymbol{v}) := \arg\min_{\boldsymbol{x}\in\mathbb{R}^m}\|\boldsymbol{x} - \boldsymbol{v}\|^2 \; s.t.\, \|\boldsymbol{x}\|_1 \le 1$$

UNIVERSITY AT ALBANY
State University of New York

Experiment

Language Generation

We tested MagR on Language Generation tasks. The results for LLaMA family are shown in the following tables. MagR consistently improves the performance of RTN and OPTQ. Moreover, MagR+OPTQ outperforms most methods across the LLaMA family models for both per-channel and per-group weight quantization.

Datasets		Wikitext2			C4		
LLaMA / PPL↓		2-7B	2-13B	2-70B	2-7B	2-13B	2-70B
FP16	Baseline	5.47	4.88	3.31	6.97	6.46	5.52
W2A16	OPTQ	7.7e3	2.1e3	77.95	NAN	323.12	48.82
	OmniQuant	37.37	17.21	7.81	90.64	26.76	12.28
	QuIP	27.13	**10.09**	6.33	31.33	**13.13**	8.94
	MagR+OPTQ†	**16.73**	11.14	**5.95**	**23.73**	14.45	**8.53**
W2A16 g128	OPTQ	36.77	28.14	-	33.70	20.97	-
	OmniQuant	11.06	8.26	6.55	15.02	11.05	8.52
	MagR+OPTQ	**9.94**	**7.63**	**5.52**	**14.08**	**10.57**	**8.05**
W3A16	RTN	539.48	10.68	7.52	402.35	12.51	10.02
	OPTQ	8.37	6.44	4.82	9.81	8.02	6.57
	AWQ	24.00	10.45	-	23.85	13.07	-
	OmniQuant	6.58	5.58	3.92	8.65	7.44	6.06
	QuIP	6.50	**5.34**	3.85	8.74	7.34	6.14
	MagR+RTN	8.66	6.55	4.64	10.78	8.26	6.77
	MagR+OPTQ	**6.41**	5.41	**3.82**	**8.23**	**7.19**	**6.03**
W3A16 g128	RTN	6.66	5.51	3.97	8.40	7.18	6.02
	OPTQ	6.29	5.42	3.85	7.89	7.00	5.85
	AWQ	6.24	5.32	-	7.84	6.94	-
	OmniQuant	6.03	5.28	3.78	**7.75**	6.98	5.85
	MagR+RTN	6.46	5.45	3.95	8.22	7.12	6.00
	MagR+OPTQ	**6.00**	**5.23**	**3.71**	7.77	**6.93**	**5.84**
W4A16	RTN	6.11	5.20	3.67	7.71	6.83	5.79
	OPTQ	5.83	5.13	3.58	7.37	6.70	5.67
	AWQ	6.15	5.12	-	7.68	6.74	-
	OmniQuant	5.74	5.02	3.47	7.35	6.65	5.65
	QuIP	5.94	5.01	3.53	8.01	6.88	5.87
	MagR+RTN	5.91	5.17	3.58	7.52	6.81	5.72
	MagR+OPTQ	**5.70**	**4.97**	**3.44**	**7.28**	**6.63**	**5.63**

Datasets		Wikitext2				C4			
LLaMA / PPL↓		1-7B	1-13B	1-30B	1-65B	1-7B	1-13B	1-30B	1-65B
FP16		5.68	5.09	4.10	3.53	7.08	6.61	5.98	5.62
W2A16	OPTQ	2.1e3	5.5e3	499.75	55.91	689.13	2.5e3	169.80	40.58
	OmniQuant	**15.47**	13.21	8.71	7.58	24.89	18.31	13.89	10.77
	MagR+OPTQ†	19.98	**9.41**	**8.47**	**6.41**	**24.69**	**16.37**	**13.09**	**8.82**
W2A16 g128	OPTQ	44.01	15.60	10.92	9.51	27.71	15.29	11.93	11.99
	OmniQuant	**9.72**	**7.93**	7.12	**5.95**	**12.97**	**10.36**	9.36	**8.00**
	MagR+OPTQ	9.89	9.22	**6.72**	6.41	13.14	10.62	**8.05**	9.14
W3A16	RTN	25.73	11.39	14.95	10.68	28.26	13.22	28.66	12.79
	OPTQ	8.06	6.76	5.84	5.06	9.49	8.16	7.29	6.71
	AWQ	11.88	7.45	10.07	5.21	13.26	9.13	12.67	7.11
	OmniQuant	**6.49**	5.68	4.74	**4.04**	**8.19**	7.32	6.57	**6.07**
	MagR+RTN	7.93	6.71	5.66	4.79	9.77	8.46	7.38	6.87
	MagR+OPTQ	6.86	**5.43**	**4.73**	4.2	8.65	**7.21**	**6.56**	6.16
W3A16 g128	RTN	7.01	5.88	4.87	4.24	8.62	7.49	6.58	6.10
	OPTQ	6.55	5.62	4.80	4.17	7.85	7.10	6.47	6.00
	AWQ	6.46	5.51	4.63	3.99	7.92	7.07	6.37	5.94
	OmniQuant	**6.15**	5.44	4.56	**3.94**	**7.75**	**7.05**	**6.37**	5.93
	MagR+RTN	6.90	5.50	4.82	4.17	8.46	7.19	6.52	6.02
	MagR+OPTQ	6.29	**5.41**	**4.52**	3.95	7.78	7.09	6.38	**5.93**
W4A16	RTN	6.43	5.55	4.57	3.87	7.93	6.98	6.34	5.85
	OPTQ	6.13	5.40	4.48	3.83	7.43	6.84	6.20	5.80
	AWQ	6.08	5.34	4.39	3.76	7.52	6.86	6.17	5.77
	OmniQuant	**5.86**	**5.21**	4.25	**3.71**	**7.34**	**6.76**	**6.11**	**5.73**
	MagR+RTN	6.16	5.42	4.36	3.80	7.66	6.87	6.22	5.82
	MagR+OPTQ	6.03	5.23	**4.24**	3.72	7.39	6.77	6.13	5.75

UNIVERSITY AT ALBANY
State University of New York

Zero-shot tasks

We evaluated the performance of quantized models on four zero-shot tasks, shown in the following table.

LLaMA2 / Acc↑	WBits	Method	ARC-C	ARC-E	PIQA	Winogrande	Avg.
LLaMA2-7B	FP16	-	40.0	69.3	78.5	67.3	63.8
	4	OmniQuant	37.9	67.8	77.1	**67.0**	62.5
	4	MagR+OPTQ	**39.3**	**68.4**	**78**	66.5	**63.1**
	3	OmniQuant	**35.3**	**62.6**	73.6	**63.6**	**58.8**
	3	MagR+OPTQ	34.6	62	**74.7**	63	58.6
	2	OmniQuant	21.6	35.2	57.5	**51.5**	41.5
	2	QuIP	19.4	26.0	54.6	51.8	37.5
	2	MagR+OPTQ†	**22.0**	**36.7**	**59.8**	51.1	**42.4**
LLaMA2-13B	FP16	-	45.6	73.3	79.1	69.6	66.9
	4	OmniQuant	43.1	70.2	78.4	67.8	64.9
	4	QuIP	**44.9**	**73.3**	**79**	**69.7**	**66.7**
	4	MagR+OPTQ	44.2	72.0	78.0	68.6	65.7
	3	OmniQuant	42.0	69.0	77.7	65.9	63.7
	3	QuIP	41.5	**70.4**	76.9	**69.9**	**64.7**
	3	MagR+OPTQ	**42.2**	69.0	**77.7**	66.5	63.9
	2	OmniQuant	23.0	44.4	**62.6**	52.6	45.7
	2	QuIP	**23.5**	**45.2**	62.0	**52.8**	**45.9**
	2	MagR+OPTQ†	23.2	44.3	62.4	52.1	45.5
LLaMA2-70B	FP16	-	51.1	77.7	81.1	77.0	71.7
	4	OmniQuant	49.8	**77.9**	80.7	75.8	71.1
	4	QuIP	47.0	74.3	80.3	76.0	69.4
	4	MagR+OPTQ	**50.1**	77.5	**80.8**	**76.0**	**71.1**
	3	OmniQuant	47.6	75.7	79.7	73.5	69.1
	3	QuIP	46.3	73.2	**80.0**	74.6	68.5
	3	MagR+OPTQ	**47.7**	**76.6**	79.4	**75.4**	**69.8**
	2	OmniQuant	28.7	55.4	68.8	53.2	51.5
	2	QuIP	34.0	**62.2**	**74.8**	**67.5**	**59.6**
	2	MagR+OPTQ†	**35.9**	61.3	74.7	64.8	59.2

The adaptive capacity of MagR

We investigated the combined effects of MagR and QuIP. MagR significantly enhances the performance of QuIP on LLaMA2 models family.

Datasets		Wikitext2		C4	
LLaMA / PPL↓		2-7B	2-13B	2-7B	2-13B
FP16	Baseline	5.47	4.88	6.97	6.46
W2A16	QuIP	27.13	10.09	31.33	13.13
	MagR+QuIP	13.31	9.40	14.49	11.07
W3A16	QuIP	6.50	5.34	8.74	7.34
	MagR+QuIP	6.25	5.29	7.88	7.02
W4A16	QuIP	5.94	5.01	8.01	6.88
	MagR+QuIP	5.74	4.99	7.25	6.63

UNIVERSITY AT ALBANY
State University of New York

Concluding Remarks

This paper introduces MagR, a method to shrink weight sizes in pre-trained language models, helping improve accuracy in common quantization methods. By grouping weights closer together, MagR allows for smaller quantization steps. Experiments on LLaMA models show state-of-the-art performance without any inference overhead, making MagR practical for quantized model deployment.

Thank you!