AFFINEQUANT: AFFINE TRANSFORMATION QUANTIZATION FOR LARGE LANGUAGE MODELS

Yuexiao Ma[1]*, **Huixia Li**[2], **Xiawu Zheng**[1,3,4], **Feng ling**[2], **Xuefeng Xiao**[2],
Rui Wang[2], **Shilei Wen**[2], **Fei Chao**[1], **Rongrong Ji**[1,4]†
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, School of Informatics, Xiamen University, 361005, P.R. China.
[2] ByteDance Inc. [3] Peng Cheng Laboratory, Shenzhen, China.
[4] Institute of Artificial Intelligence, Xiamen University.

Background

- 大语言模型（LLM）在自然语言处理任务中展现出卓越性能，但其庞大的参数规模限制了在资源受限设备上的部署，如移动设备和边缘计算场景。研究如何高效压缩LLM，使其在保持性能的同时适应这些设备，具有重要的实际应用价值，可推动LLM在更多场景的广泛应用。
- 量化技术正是在这种“大模型瘦身”的方法之一：
  把FP16/FP32权重或激活映射到INT8、INT4甚至更低比特，可瞬间把模型体积缩小2-8倍，显存带宽和计算能耗同步下降；若配合整数矩阵乘（INT GEMM）或专用指令集，推理速度还能再提1.5-4倍。

Question & Idea

- 但现有 PTQ（如 AWQ、OmniQuant）仅支持缩放（scale）或平移（shift）等简单等价变换，限制了优化空间，导致低比特（尤其 W4A4、W3A16）下误差大。
- 本文介绍了AffineQuant，一种利用等价仿射变换进行量化的方法。与其他方法相比，AffineQuant在整个优化过程中始终保持最优的均方误差。
- 本文还探究了仿射变换矩阵在优化过程中的可逆性。为了保证稳定性，提出了一种基于Levy -德普朗克定理的渐进掩码方法来保持仿射变换矩阵为严格对角占优矩阵。

Contributions

- We propose a novel affine transform in PTQ, which retains the benefits of PTQ, assures efficiency and generalization, significantly minimizes quantization error, especially under low-bit quantization, and enables the deployment of LLMs on edge devices.
- We propose a novel optimization algorithm that guarantees invertibility throughout the process, utilizing the Levy-Desplanques theorem, and simultaneously reduces computational costs.
- Our method obtains the state-of-the-art performance for large language model quantization, especially on low-bit or small models. Without additional overhead, on the w4a4 configuration of LLaMA2-7B, our perplexity on the C4 dataset is 15.76 (2.26↓ vs 18.02 in OmniQuant). Similarly, on the w4a4 configuration of LLaMA-30B, our accuracy on 6 zero-shot tasks is 58.61% (1.98 ↑ vs 56.63 in OmniQuant).

We define the pseudo-quantization function as follows:

$$\mathcal{Q}(x) = \Delta * \left(clamp\left(\left\lfloor \frac{x}{\Delta} \right\rceil + zp, 0, 2^n - 1 \right) - zp \right), \tag{1}$$

where Δ, zp and n are the quantization step-size, zero point and bits, respectively. $\lfloor \cdot \rceil$ is the rounding operation. As depicted in Figure 1, AffineQuant involves left-multiplying the affine transform matrix A by weight matrix W to better align the weight distribution with the quantization function $Q(\cdot)$. Expanding the optimization space enables smaller quantization errors in the transformed weights, leading to a reduction in perplexity. Simultaneously, we right-multiply the inverse of the affine transform matrix A by the activation value X to maintain the invariance of the matrix multiplication output between activations and weights. For a single linear layer, AffineQuant formulates the following optimization problem:

$$\arg\min_{\boldsymbol{A}} \left\| XW - XA^{-1}\mathcal{Q}(AW) \right\|_F^2. \tag{2}$$

$$PPL \propto \mathcal{L}_{CE} \propto \left\| XW - XA^{-1}\mathcal{Q}(AW) \right\|_F^2, \tag{3}$$

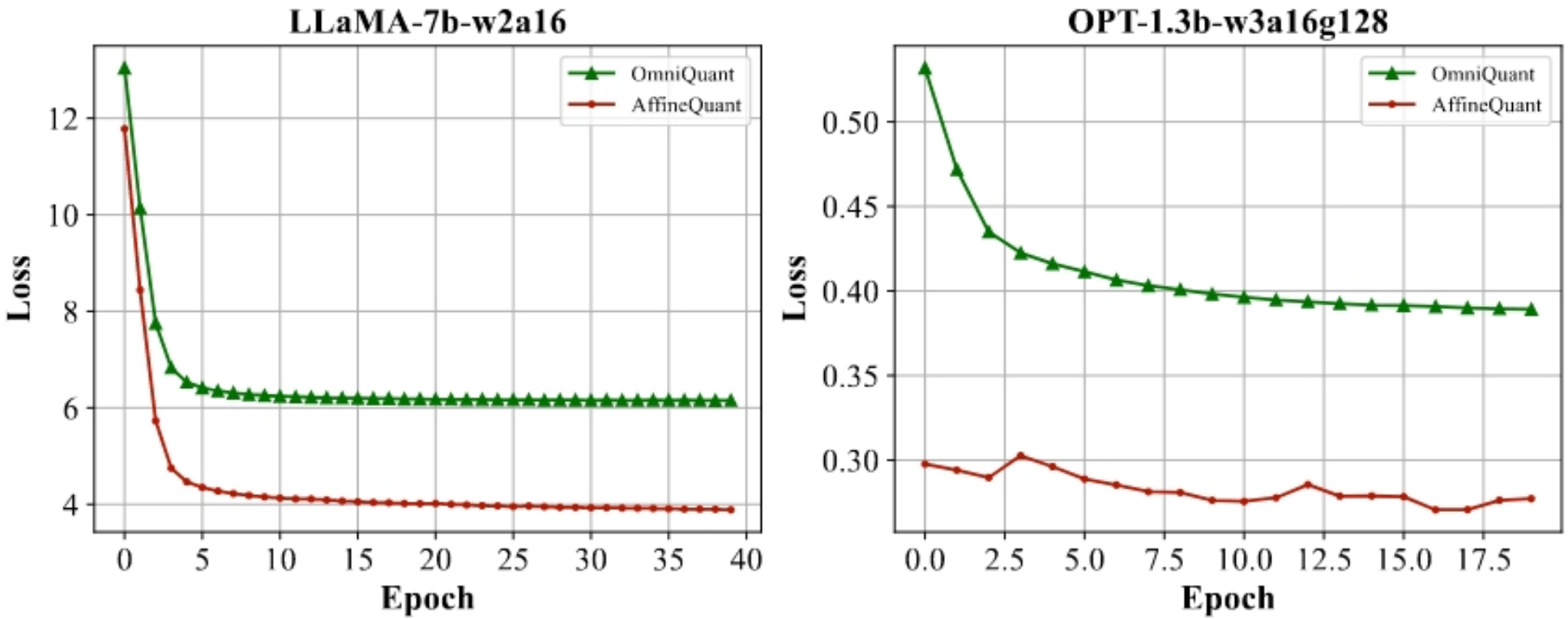


Figure 3: Mean square error loss of the last transformer block of LLaMA-7b and OPT-1.3b. "w2a16" means 2-bit weight-only quantization. "w3a16g128" means 3-bit grouping 128 weight-only quantization. We optimize 40 and 20 epochs in the last block of LLaMA-7b and OPT-1.3b, respectively.

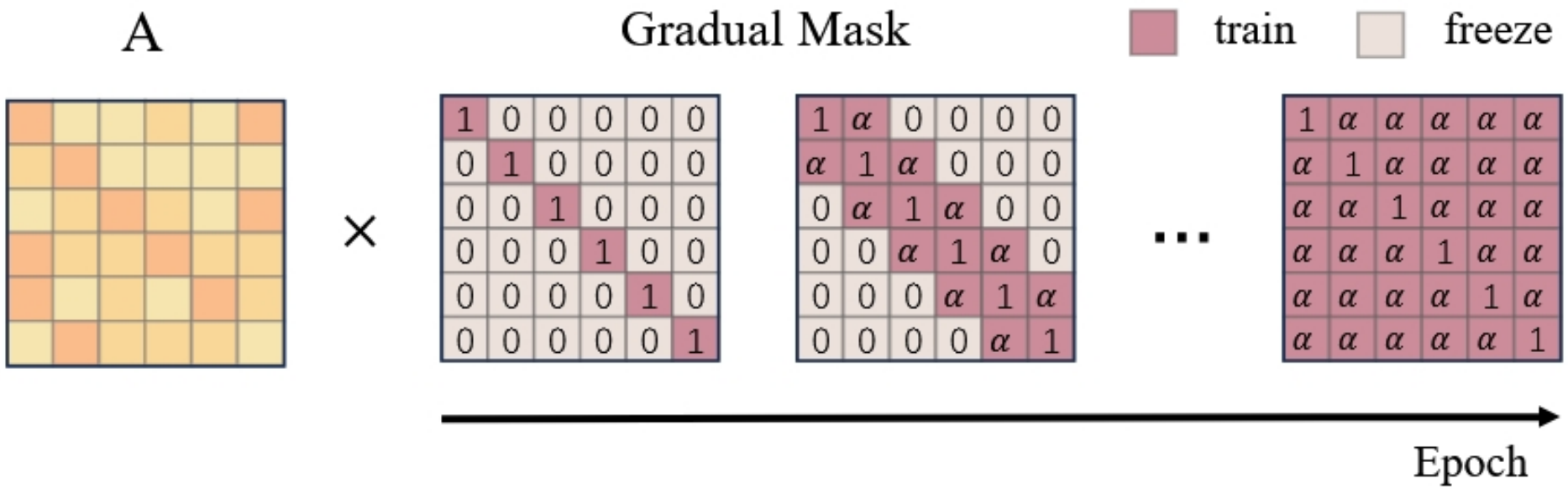


Figure 2: The gradual mask operates on the affine transformation matrix, gradually incorporating the elements of matrix A near the diagonal into the training process as training progresses.

$$|a_{ii}| > \sum_{i \neq j} |a_{ij}|, \quad for\ all\ i. \qquad GM_{ij} = \begin{cases} 1 & i = j, \\ \alpha & 0 < |i-j| \leq \frac{e}{t} \times hidden\ size, \\ 0 & otherwise, \end{cases}$$

Experiment

Table 1: Weight-only quantization PPL(↓) results on the OPT model WikiText2 dataset.

Config	Method	125M	1.3B	2.7B	6.7B	13B	30B
FP16	-	27.65	14.63	12.47	10.86	10.12	9.56
w3a16	RTN	1.2e3	1.3e4	1.6e4	6.5e3	4.6e3	1.5e3
	GPTQ (Frantar et al., 2022)	53.05	21.17	16.83	15.09	11.73	10.30
	AWQ (Lin et al., 2023)	69.43	28.01	263.10	15.13	20.09	35.74
	OmniQuant (Shao et al., 2023)	35.66	16.68	13.80	11.65	10.87	10.00
	AffineQuant	30.56	15.94	13.15	11.44	10.76	9.98
w3a16g128	RTN	51.22	119.00	297.98	23.54	46.03	18.80
	GPTQ (Frantar et al., 2022)	39.24	16.47	13.69	11.65	10.35	9.73
	AWQ (Lin et al., 2023)	36.74	16.32	13.58	11.41	10.68	9.85
	OmniQuant (Shao et al., 2023)	32.25	15.72	13.18	11.27	10.47	9.79
	AffineQuant	30.21	15.61	12.98	11.18	10.51	9.81
w4a16	RTN	37.28	48.17	16.92	12.10	11.32	10.97
	GPTQ (Frantar et al., 2022)	31.43	15.56	12.82	11.41	10.31	9.63
	AWQ (Lin et al., 2023)	32.28	15.49	12.93	11.30	10.39	9.77
	OmniQuant (Shao et al., 2023)	29.45	15.04	12.76	11.03	10.30	9.65
	AffineQuant	28.39	14.92	12.64	10.96	10.26	9.65
w4a16g128	RTN	30.47	15.29	13.02	11.15	10.30	9.94
	GPTQ (Frantar et al., 2022)	29.81	14.89	12.52	10.93	10.17	9.58
	AWQ (Lin et al., 2023)	29.15	14.94	12.74	10.93	10.21	9.59
	OmniQuant (Shao et al., 2023)	28.86	14.88	12.65	10.96	10.20	9.62
	AffineQuant	28.33	14.79	12.58	10.92	10.19	9.62

Experiment

Table 2: AffineQuant and OmniQuant quantization performance of LLaMA-7B, 13B, 30B on six zero-shot datasets using 4/4 bit quantization.

	Dataset	PIQA (↑)	ARC-e (↑)	WinoGrande (↑)	BoolQ (↑)	ARC-c (↑)	HellaSwag (↑)	Avg. (↑)
LLaMA-7B w4a4	FP16	77.47	52.48	67.07	73.08	41.46	73.00	64.09
	OmniQuant Shao et al. (2023)	66.15	45.20	53.43	63.51	31.14	56.44	52.65
	AffineQuant	69.37	42.55	55.33	63.73	31.91	57.65	53.42
LLaMA-13B w4a4	FP16	79.10	59.89	70.31	68.01	44.45	76.21	66.33
	OmniQuant Shao et al. (2023)	69.69	47.39	55.80	62.84	33.10	58.96	54.37
	AffineQuant	66.32	43.90	54.70	64.10	29.61	56.88	52.58
LLaMA-30B w4a4	FP16	80.08	58.92	72.53	68.44	45.47	79.21	67.44
	OmniQuant Shao et al. (2023)	71.21	49.45	59.19	65.33	34.47	64.65	56.63
	AffineQuant	70.84	49.41	58.64	70.12	37.12	65.53	58.61

Experiment

Table 3: Quantization performance of LLaMA1&2 on WikiText2 and C4 datasets using 4/4 bit weight-activation quantization.

Datasets	LLaMA1&2	Methods	1-7B	1-13B	1-30B	2-7B	2-13B
WikiText2	FP16	-	5.68	5.09	4.10	5.47	4.88
	W4A4	SmoothQuant Xiao et al. (2023)	25.25	40.05	192.40	83.12	35.88
		OmniQuant Shao et al. (2023)	11.26	10.87	10.33	14.26	12.30
		AffineQuant	10.28	10.32	9.35	12.69	11.45
C4	FP16	-	7.08	6.61	5.98	6.97	6.46
	W4A4	SmoothQuant Xiao et al. (2023)	32.32	47.18	122.38	77.27	43.19
		OmniQuant Shao et al. (2023)	14.51	13.78	12.49	18.02	14.55
		AffineQuant	13.64	13.44	11.58	15.76	13.97

Ablation Study

Table 5: Effect of different stability factors α on model performance of OPT-125M and LLaMA-7B.

	Dataset	FP16	1e0	1e−1	1e−2	1e−3	1e−4	1e−5	1e−6	1e−7	1e−8
OPT-125M w2a16g128	WikiText2	27.65	42.08	45.32	65.77	76.03	76.78	76.00	76.48	76.55	76.46
	PTB	32.54	63.60	68.29	114.41	135.72	125.38	124.12	132.36	132.45	113.75
	C4	24.60	45.80	50.84	70.44	79.60	81.48	79.72	79.67	80.70	79.54
LLaMA-7B w2a16	WikiText2	5.68	NaN	NaN	9.53	10.63	10.90	10.99	10.72	11.63	11.67
	C4	7.08	NaN	NaN	14.89	14.62	15.31	15.22	14.78	16.66	17.33

Table 6: Contributions of gradual mask on OPT-125M and LLaMA-7B model.

	Scheme	WikiText2	PTB	C4		Scheme	WikiText2	C4
OPT-125M w3a16	FP16	27.65	32.54	24.60	LLaMA-7B w2a16	FP16	5.68	7.08
	With Gradual	32.10	39.85	29.97		With Gradual	9.53	14.89
	Without Gradual	53.52	90.47	62.17		Without Gradual	NaN	NaN

Contribution

- 提出的渐进掩码保证了矩阵在优化过程中保持严格对角占优，保证了矩阵的可逆性，并进一步降低了均方误差目标函数。
- 提出了一种新的仿射变换，它保留了PTQ的优点，保证了效率和泛化性，显著地减少了量化误差，特别是在低比特量化的情况下，并允许在边缘设备上部署LLM。
- 将仿射变换矩阵与其他层进行整合，即将仿射变换矩阵与权重和偏置参数进行合并，不引入任何额外的开销，降低了计算成本。