LO-BCQ: Locally Optimal Block Clustered Quantization

Reena Elangovan[1], Charbel Sakr[1], Anand Raghunathan[2], Brucek Khailany[1]

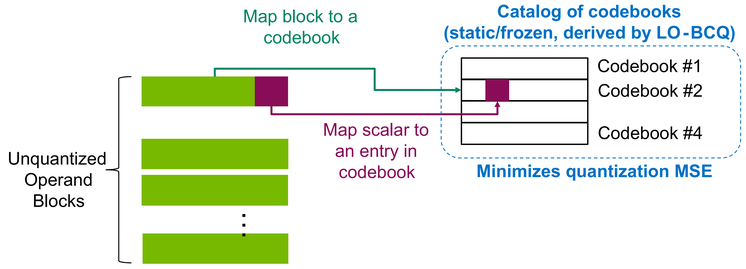
[1] NVIDIA

[2] PURDUE UNIVERSITY

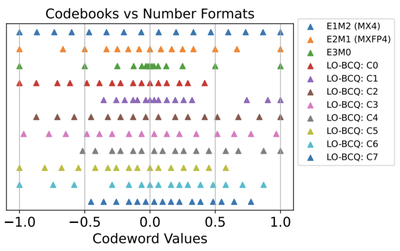
MSE-optimal algorithm to find a catalog of scalar codebooks (LUTs) – achieves SOTA accuracy during W4A4 quantization

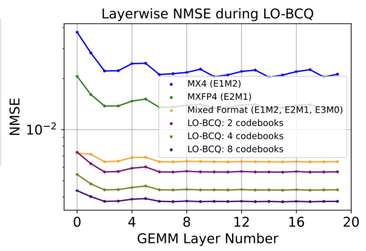
Introduction

Hybrid of scalar and vector quantization



- Group of vectors map to a scalar codebook (LUT)
- 2-level look-up for decode
- Static catalog of codebooks calibrated offline, each with INT6 entries, $\leq$ 16 codebooks
- **Cheap encoding and decoding:** Applicable to both weight and activation quantization





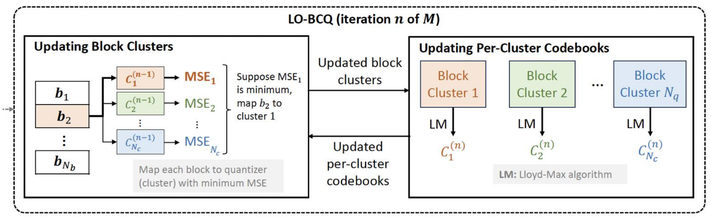
- The codebooks identified by LO-BCQ **effectively represent non-uniform value distributions** that are difficult to represent using fixed formats like FP, INT, MXFP etc.
- Better NMSE compared to hand-picked mixture of formats

Method

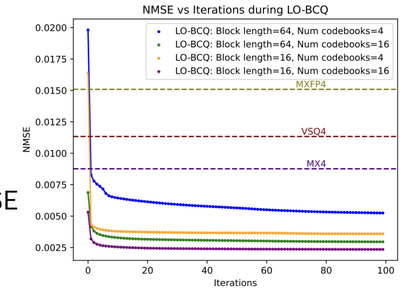
Step 0: Start with initial set of codebooks

Repeat until convergence:

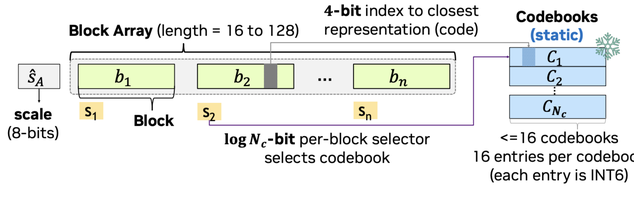
1. Cluster operand blocks by mapping to codebook that best represents them
2. Update codebook by applying Lloyd-Max on its block cluster



- Convergence is typically achieved within **100 iterations**
- More codebooks in catalog -> lower NMSE



Data-Format



- Each block is mapped to a codebook. The mapping is stored as a per-block selector
- Scale-factor is shared by a group of blocks called block array

Results

LO-BCQ achieves <0.1 loss in Wiki PPL and <1% loss on LM-Harness tasks during W4A4 quantization

Method (g128)	Bitwidth (W4A4)	Δ Wiki PPL Llama2-7B	Δ Wiki PPL Llama2-70B
SmoothQuant	4.13	77.65	-
OmniQuant	4.13	9.14	-
QuaRot	4.13	0.46	0.29
Atom	4.13	0.56	0.36
LO-BCQ ($N_c = 2$)	4.19	0.14	**0.09**
LO-BCQ ($N_c = 4$)	4.31	0.12	**0.07**
LO-BCQ ($N_c = 8$)	4.44	**0.09**	**0.06**
LO-BCQ ($N_c = 16$)	4.56	**0.08**	**0.05**

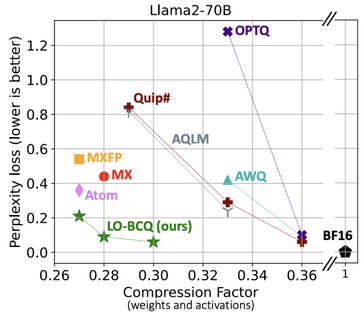
		Llama2-70B					
BF16	16	48.8	85.23	79.95	81.56	65.27	72.16
QuaRot (g128)	4.13	-	-	76.24	82.43	-	-
MX4 (g16)	4.5	**48.04**	82.41	76.40	**80.58**	63.24	70.13 (2.03)
VSQ (g16)	4.5	**47.85**	82.29	77.27	79.82	61.40	69.73 (2.43)
MXFP4 (g32)	4.25	47.75	83.06	76.32	**80.58**	63.24	70.19 (1.97)
LO-BCQ (g64, $N_c = 2$)	4.25	**49.0**	82.82	78.77	**81.45**	64.21	**71.25 (0.91)**
LO-BCQ (g64, $N_c = 8$)	4.5	**49.28**	84.03	78.37	**81.45**	**64.76**	**71.58 (0.58)**
LO-BCQ (g32, $N_c = 16$)	4.75	**49.28**	**84.93**	**80.66**	**81.34**	**65.18**	**72.28 (+0.12)**

2-bit and 3-bit Weight Quantization

Method		Bitwidth	#codebooks	Wiki2 PPL Llama2-7B	Wiki2 PPL Llama2-13B	Wiki2 PPL Llama2-70B
BF16 (Pretrained)		16	-	5.47	4.88	3.32
W3A16	QuIP# (LDLQ, no FT)	3	$2^{16} + 2^8$	5.91	5.23	3.61
	AQLM (FT)	3	$2 * 2^{12}$	5.46	4.82	3.36
	LO-BCQ (LDLQ, no FT)	3.375	4	5.79	5.12	3.53
		3.5	8	5.72	5.09	3.49
W2A16	QuIP# (LDLQ, no FT)	2	2^{16}	8.05	6.59	4.44
	AQLM (FT)	2	2^{16}	6.59	5.60	3.94
	LO-BCQ (LDLQ, no FT)	2.375	4	8.02	7.12	4.52
		2.5	8	6.87	6.16	4.20

When combined with LDLQ, *LO-BCQ **with $\leq$ 8 codebooks** achieves **lower perplexity than QuIP#***

Conclusion



- LO-BCQ achieves **SOTA PTQ accuracy** during W4A4 quantization.
- Requires small number of **static codebooks (<= 0.19KB)**