Matryoshka Quantization

Pranav Nair*, **Puranjay Datta***, Jeff Dean, Prateek Jain, Aditya Kusupati

Google DeepMind

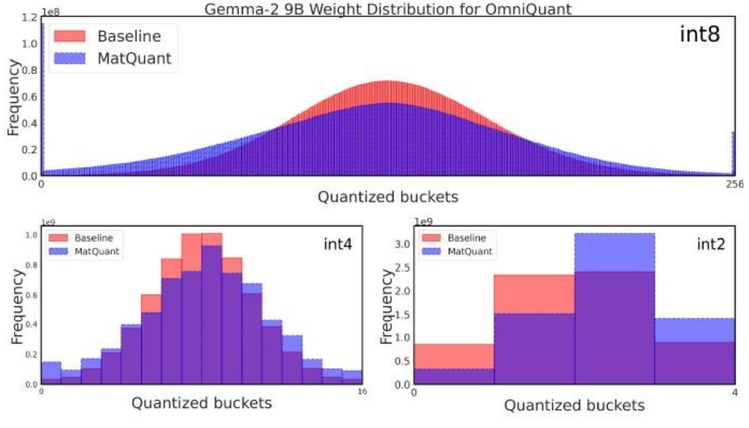
ICML International Conference On Machine Learning

Motivation

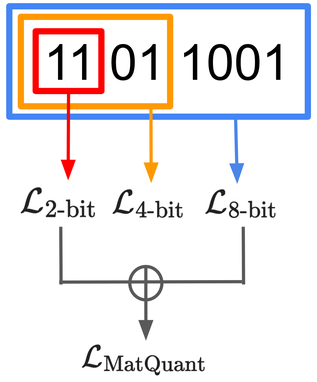
- MatQuant **jointly optimizes** for several precisions **(int8, int4, and int2)** at once yields a single model that does well on all the bit-widths it was trained on.
- **Interpolated** precisions **int6 and int3 ~ baselines**.
- **Layer-wise Mix'n'Match** with different precisions at different layers **without explicit training**, yielding combinatorially many models **free of cost**.
- Sliced **int2 MatQuant sub model >> baseline int2** model.

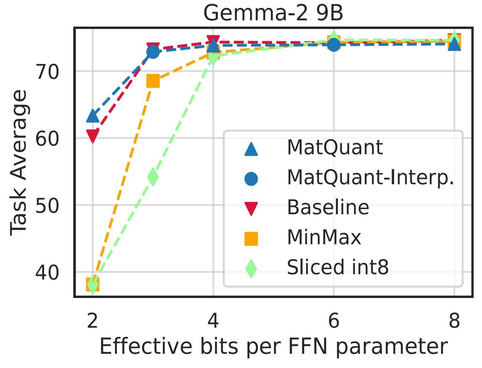
Weight Distribution

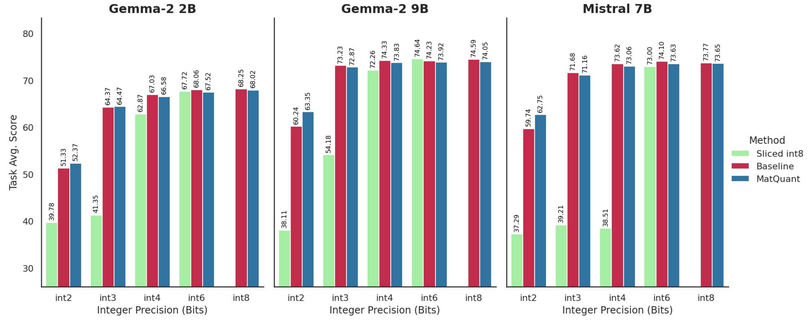
- MatQuant shifts the quantized weight distribution toward higher values, contributing to improved int2 performance.
- Int8's overparameterization provides flexibility accommodating int2's high-valued weight needs during joint training.



MatQuant



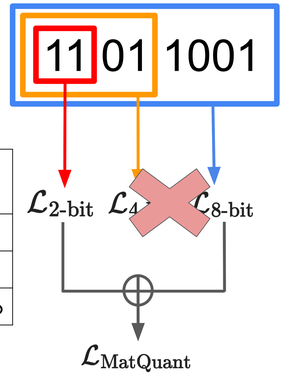




Single Precision MatQuant

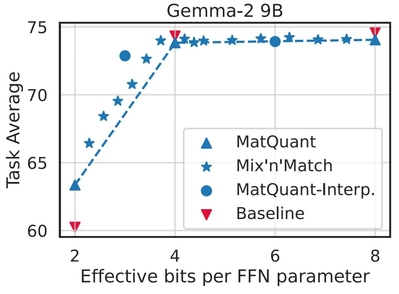
- Stabilises FFN + Attn quantization
- Improves int2 performance

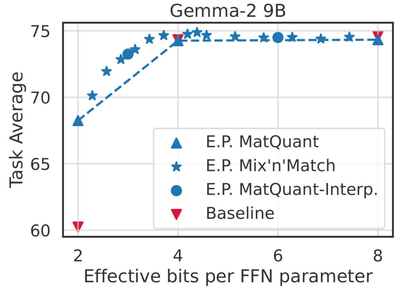
Data Type	Gemma-2 9B OmniQuant	Task Avg.	Log pplx.
INT2	S.P. MatQuant	**64.02**	**3.171**
	Baseline	60.24	3.292
	Sliced Int8	38.11	15.226



Mix'n'Match

- Densely spans the Pareto-optimal accuracy-vs-bits-per-FFN-parameter
- Combinatorial number models for free without any explicit training
- Pyramidal works for the best: 222....444....888....444....222





Codistillation

- Higher-precision model outputs as targets for lower-precision nested models

Data Type	Gemma-2 9B		OmniQuant		QAT	
			Task Avg.	Log pplx.	Task Avg.	Log pplx.
INT2	MatQuant	[8, 4, 2]	63.35	3.187	62.29	2.660
		[8, 4, 8 -> 2]	62.64	3.289	62.31	2.670
		[8, 4, 2, 8 -> 2]	62.91	3.138	**62.70**	2.673
		[8, 4, 2, 8 -> 4; 2]	**64.32**	3.227	62.60	2.670
	Baseline		60.24	3.292	56.02	2.923

Extra Precision MatQuant

- 2^r + 1 possible values instead of 2^r for an r-bit model
- <1% of weights occupying the extra bucket, yields a 5% performance boost.

Data Type	Gemma-2 9B OmniQuant		Task Avg.	Log pplx.
INT2	E.P. MatQuant	[8, 4, 2]	68.52	2.809
		[8, 4, 8 -> 2]	69.2	2.796
		[8, 4, 2, 8 -> 2]	70.17	2.778
		[8, 4, 2, 8 -> 4; 2]	**69.72**	2.804