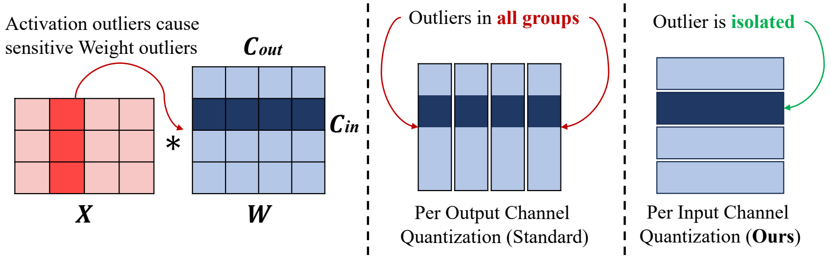
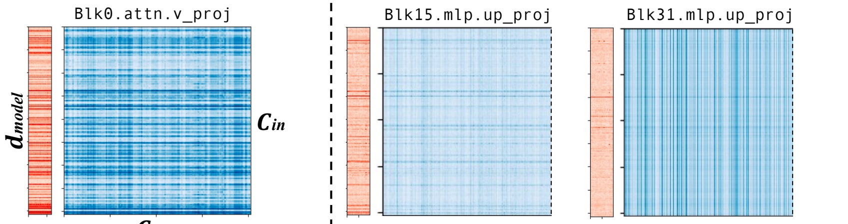
Rethinking Channel Dimensions to Isolate Outliers for Low-bit Weight Quantization of Large Language Models

Jung Hwan Heo*[1], Jeonghoon Kim*[2], Beomseok Kwon[2], Byeongwook Kim[2], Se Jung Kwon[2], Dongsoo Lee[2]
*Equal Contribution, [1]University of Southern California, [2]NAVER Cloud

Standard quantization settings have *pervasive* outliers
Our per-IC-quantization method *isolates* outliers



Outlier patterns occur in both channel dimensions in a 2D weight matrix (dominant rows or columns)



Outlier activations → Row (IC) sensitivity

Mild activations → Less sensitive rows; dominant sensitivity dimension (IC or OC) can change across network depth

Adaptive Dimensions (AdaDim): Adaptive channel quant. via *selective* and *automatic* application of Per-IC-quant.

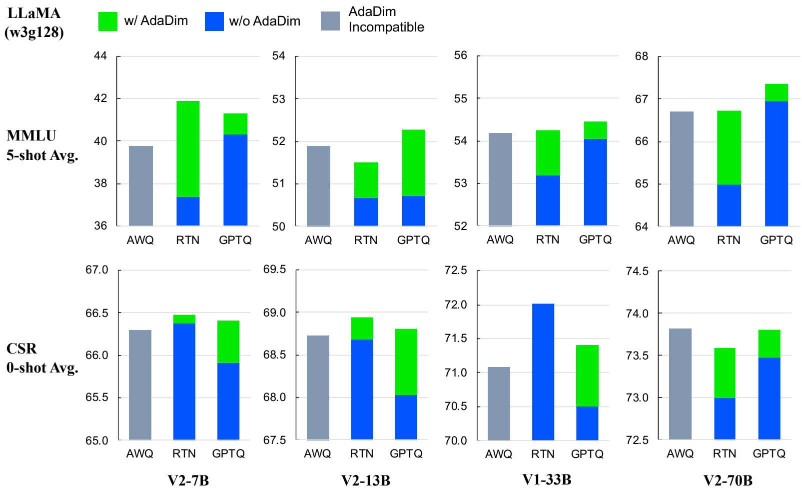
Model Size	Metric	FP16	Baseline (RTN, *Per-OC*)	(1)attn.qkv	(2)mlp.down	(1)&(2)	All
				Module to apply *Per-IC* quant.			
7B	Wiki-2 ppl. (↓)	8.79	9.22	9.17	9.11	**9.09**	9.11
	MMLU 5-shot (↑)	45.98	44.54	44.77	44.7	**45.21**	44.38
13B	Wiki-2 ppl. (↓)	7.89	8.13	8.12	8.11	**8.10**	8.13
	MMLU 5-shot (↑)	55.61	54.43	54.76	54.90	**54.97**	54.67

Simple binary selection of quant. dimension by using the reconstruction error metric

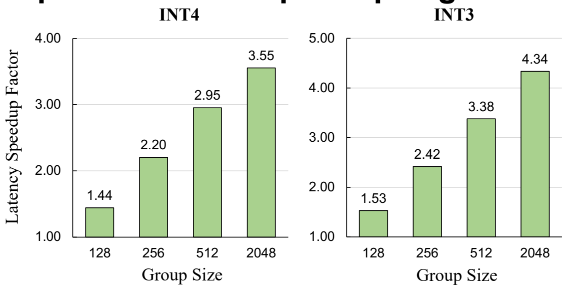
$$\mathbf{dim}^* = \underset{\mathbf{dim}\in\{oc,ic\}}{\arg\min}\ \mathcal{L}(\mathbf{dim}), \quad \mathcal{L}(\mathbf{dim}) = \|Q_{\mathbf{dim}}(\mathbf{W}_\ell)\mathbf{X}_\ell - \mathbf{W}_\ell\mathbf{X}_\ell\|,$$

Results

- **AdaDim significantly improves RTN and GPTQ for reasoning and knowledge of base LLMs**



- **Per-IC-quant. leads to speedups against CuBLAS**

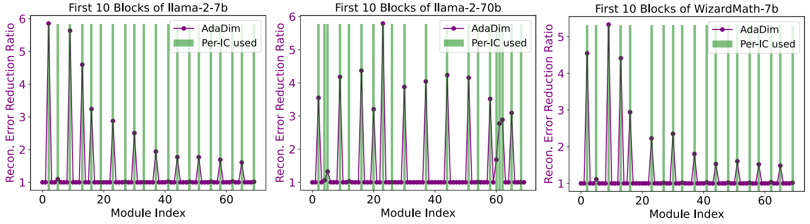


Method	Group Size	Weight ($4m \times m$)	Latency [ms]
cuBLAS		INT4	0.7258
OPTQ	128	INT3	0.3599
AWQ	128	INT4	0.3238
LUT-GEMM	128	INT4	0.2688
LUT-GEMM	128	INT3	0.225
per-IC kernel (Ours)	128	INT4	0.50477
	256	INT4	0.32926
	512	INT4	0.24570
	2048	INT4	0.20425

- **Task-specific Quantization for Math and Coding**

w3g128	GSM8k pass@1 (↑)				HumanEval pass@1 (↑)			
	WizMath-7B		WizMath-13B		WizCoder-Py-7B		WizCoder-Py-13B	
FP16	55.35		63.38		55.49		64.02	
RTN	32.52		49.13		35.37		50.61	
calib. set	base	target	base	target	base	target	base	target
AWQ	39.42	40.49	55.19	54.97	43.29	44.51	57.32	60.36
RTN-ada	37.38	39.12	50.95	53.15	42.68	45.12	**60.37**	60.98
GPTQ	38.29	41.09	54.21	57.16	31.71	45.12	53.05	56.71
GPTQ-ada	**41.77**	**42.15**	**56.78**	**57.47**	**46.34**	**46.95**	53.69	**62.2**

- **AdaDim lowers reconstruction error up to 6x**



References

- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. "Awq: Activation-aware weight quantization for llm compression and acceleration." *MLSys 2024*
- Frantar, Elias, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. "Gptq: Accurate post-training quantization for generative pre-trained transformers." *ICLR 2023*
- Kim, Sehoon, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. "Squeezellm: Dense-and-sparse quantization." *arXiv, 2023*
- Park, Gunho, Baeseong Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, and Dongsoo Lee. "Lut-gemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models." *ICLR 2024*

> *Code available at: github.com/johnheo/adadim-llm*