Efficient Quantization of Mixture-of-Experts with Theoretical Generalization Guarantees

Mohammed Nowaz Rabbani Chowdhury[1], Kaoutar El Maghraoui[2], Hsinyu Tsai[2], Naigang Wang[2], Geoffrey W. Burr[2], Liu Liu[1], Meng Wang[1]

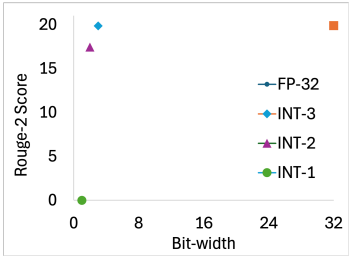
[1]Rensselaer Polytechnic Institute

[2]IBM Research

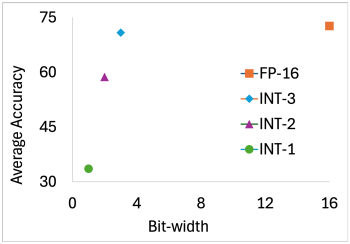
The Fourteenth International Conference on Learning Representations (ICLR 2026)
April 23-27, 2026

Background and Motivation

- Sparse MoE $\rightarrow$ Efficient scaling of deep models
- Requires huge memory for inference
- Post-training weight quantization (PTWQ) $\Rightarrow$ Promising technique for memory requirements of LLMs
- Uniform bit-width over experts [Frantar et. al. 24] $\Rightarrow$ Significant performance degradation in extremely low-bit settings



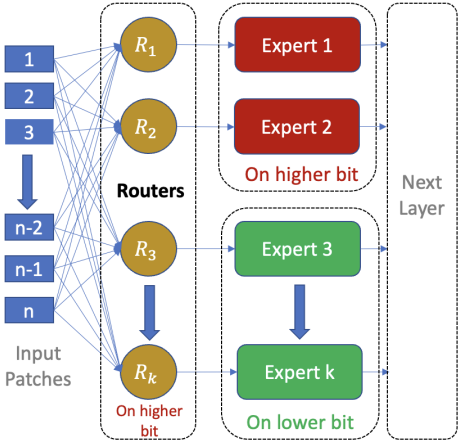
Switch Transformer on summarization task



Mixtral 8x7B on eight benchmark LLM tasks

Expert-wise Mixed-precision Quantization of MoE

- Varying importance/ sensitivity of the experts $\Rightarrow$ Expert-wise mixed-precision Quantization
- How to categorize experts in different groups of bit-width?
- Previous works [Li et. al. 24, Huang et. al. 25]:
  - Calibration data-depended, heuristic methods
  - Overlook the varying sensitivity of model performance to the quantization of different experts
  - Substantial compute for bit-allocation
  - No theoretical performance guarantee



Expert-wise mixed-precision MoE quantization

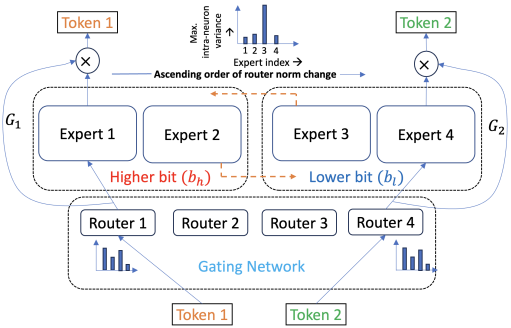
Contributions

- Proposed an expert-wise mixed-precision quantization of MoE:
  - Theoretically provable generalization guarantee
  - Calibration data-free
  - Considers both quantization noise and model's performance sensitivity
  - Insignificant overhead for bit allocation
- Major contributions:
  - A theoretically-grounded strategy, providing insights about why and how we can vary bit-width across experts
  - Superior performance over other expert-wise and non-expert-wise mixed-precision baselines
  - Reduces the inference computation compared to prior methods, and incurs negligible computational overhead

Our Method

In each MoE layer:

- Step 1: Rank experts in increasing order of routers' change of l_2 norm during training
- Step 2: Reorder lower rank experts with significantly large maximum intra-neuron variance (ζ times larger) to higher rank
- Step 3: Assign higher rank experts to higher bit



Router norm-based expert-wise mixed-precision MoE quantization

Intuition

- Large maximum intra-neuron variance $\rightarrow$ large quantization noise
- From training dynamic analysis:
  - Condition: Similar maximum intra-neuron variance for all experts
  - Lower change in router's l_2 norm during training $\rightarrow$ lower activation $\rightarrow$ model performance is more sensitive to the quantization of these experts

Setting for Theoretical Analysis

Data model:

- Binary classification
- Class 1 data contain o_1 or -o_1; Class 2 data contain o_2 or -o_2,
  - Less-prevalent feature (o_1 and o_2): appearance frequency is $\alpha < 1/4$ in the data
  - More-prevalent feature (-o_1 and -o_2): appearance frequency is $(1 - \alpha)$
- $d - 2$ task-irrelevant patterns and o_1 o_2 form an orthonormal basis.

Network model: One MoE layer; Each expert is a two-layer FFN

Theoretical Results

Theorem 1 (Expert sensitivity to noise)

The router norm change of an expert, say s_1, that learn less-prevalent features is less than that of an expert, say s_2, that learn more-prevalent features, i.e.,

$$\Delta_{s_1} < \Delta_{s_2} \tag{1}$$

The expert activation by task relevant features in s_2 is higher than that in s_1, satisfying

$$\frac{Act(s_1)}{Act(s_2)} < \frac{2\alpha}{1-2\alpha}$$

- A higher activation by a task-relevant pattern leads to a larger gap the output of this expert and the output of another expert not selecting task-relevant pattern, leading to robust predictions against quantization noise.
- Experts that learn more-prevalent patterns experience larger *change in router's l_2 norm*, and is more robust to quantization noise.

Theoretical Results

Theorem 2

Assume the max. intra-neuron var. is $\Theta(1)$ for all experts. Assign γ fraction of experts with the smaller router norm change to higher bit (b_h), rest in lower bit (b_l)

$$b_h > \log_2(1 + \Omega(d)), \text{ and } b_l > \log_2(1 + \frac{\alpha}{1-\alpha}\Omega(d))$$

then the resulting model has the same generalization accuracy as the full-precision model. Here, γ is the fraction of experts learning less-prevalent features.

Each low-precision expert can $\log_2(\frac{1-\alpha}{\alpha})$ fewer bits per value than the high-precision experts.

Empirical Results: Pre-trained MoE

Models: Mixtral 8x7B and Mixtral 8x22B [Jiang et. al. 24]

- 32 MoE layers; 8 experts/layer (Mixtral 8x7B), 56 MoE layers; 8 experts/layer (Mixtral 8x22B)
- Number of Parameters: 46.7 B (Mixtral 8x7B), 140.6 B (Mixtral 8x22B)

Tasks: 8 benchmark LLM tasks: PIQA, ARC-e, ARC-c, BoolQ, Hella Swag, Winogrande, MathQA, MMLU

Mixed-precision bit-choices: 1, 2, 3

Use the pre-trained router norm directly, no fine-turning needed.

Baselines:

- Pre-loading Mixed-precision Quantization (PMQ) [Huang et. al. 25]:
  - Computes the output error, measured by Frobenius norm, between quantizing expert s to b bits and its full-precision output, for every expert s and bit level b
  - Determines optimal bit-width assignment by minimizing the weighted sum of these errors, scaled by the expert's activation frequency and weight
- Hessian [Dong et. al. 24], BSP [Li et. al. 24] (layer-wise), and Slim-LLM [Huang et. al. 25] (group-wise)

Empirical Results: Pre-trained MoE

Our method outperforms all the baselines for Mixtral 8x7B

Model	Method	Avg. bits/exp.	Memory (GB)	Avg. Accuracy (%)
Mixtral 8x7B	Full-precision	16 (FP)	96.8	72.72
	Uniform	3	19.3	70.85
		2	13.1	58.73
	Router norm + Max var (**Ours**)	2.75	17.7	**70.01**
		2.5	16.1	**68.38**
		2.25	14.5	**65.79**
		2.0	13.1	62.56
	PMQ	2.75	17.7	69.85
		2.5	16.1	67.53
		2.25	14.5	65.16
		2.0	13.1	**62.83**
	Hessian	2.5	17.0	67.18
		2.25	15.3	63.47
		2.0	13.6	58.85
	BSP	2.5	17.0	49.07
	Slim-LLM	2.0	13.6	44.49

Empirical Results: Pre-trained MoE

Our method outperforms PMQ for Mixtral 8x22B

Model	Method	Avg. bits/exp.	Memory (GB)	Avg. Accuracy (%)
Mixtral 8x22B	Full-precision	16 (FP)	281.2	76.31
	Uniform	3	57.5	65.48
		2	38.6	36.53
	Router norm +Max var (**Ours**)	2.5	46.7	60.10
		2.25	43.0	**59.17**
		2.0	38.6	**55.34**
		1.75	35.2	**51.44**
	PMQ	2.5	46.7	**60.69**
		2.25	43.0	55.35
		2.0	38.6	54.80
		1.75	35.2	47.87

Empirical Results: Pre-trained MoE

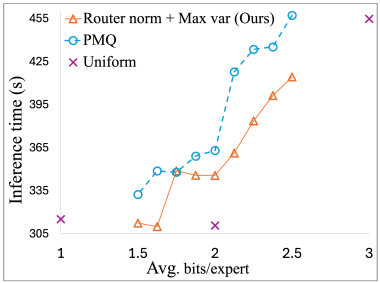
Model: Mixtral 8x7B [Jiang et. al. 24]

- 32 MoE layers, 8 experts/layer
- Number of Parameters: 46.7 billion

Task: Language modeling on WikiText-103 dataset

Remarks:

- Our method outperforms the SOTA methods (i.e., PMQ) in terms of inference speed
- Our method allocates less-frequently used experts to higher bit
- PMQ selects some more frequently used experts in higher bit



Inference efficiency

Conclusion

We propose an expert-wise mixed-precision of MoE:

- Assign higher bit to the experts with lower router norm and vice-versa
- Reorder experts with large intra-neuron variance to higher bit
- Provided theoretical generalization guarantee

Future work:

- Combining with block-wise and layer-wise mixed-precision quantization methods
- Extending to activation quantization

References I

[Chowdhury et al. ICML'23] Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen.
Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks.
In *2023 International Conference on Machine Learning (ICML)*, 2023.

[Chowdhury et al. ICML'24] Mohammed Nowaz Rabbani Chowdhury, Meng Wang, Kaoutar El Maghraoui, Naigang Wang, Pin-Yu Chen and Christopher Carothers.
A Provably Effective Method for Pruning Experts in Fine-tuned Sparse Mixture-of-Experts.
In *2024 International Conference on Machine Learning (ICML)*, 2024.

[Riquelme et al. 21] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers and N. Houlsby.
Scaling vision with sparse mixture of experts.
In *Advances in Neural Information Processing Systems*, 34, 2021.

References II

[Vaswani et al. 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin.
Attention is all you need.
In *Advances in Neural Information Processing Systems*, 30, 2017.

[Fedus et al. 22] W. Fedus, B. Zoph, and N. Shazeer. (2022).
Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity
The Journal of Machine Learning Research, 23(1):5232–5270

[Chen et al. 22)] Chen, Z., Deng, Y., Wu, Y., Gu, Q., and Li, Y.
Towards understanding mixture of experts in deep learning.
arXiv preprint arXiv:2208.02813, 2022.

References III

[Allingham et al. 22] J.U. Allingham, F. Wenzel, Z.E. Mariet, B. Mustafa, J. Puigcerver, N. Houlsby, G. Jerfel, V. Fortuin, B. Lakshminarayanan, J. Snoekand, D. Tran, C. Riquelme, and R. Jenatton.
Sparse MoEs meet Efficient Ensembles
Transactions on Machine Learning Research, 2022.

[Frantar et. al. 24] E. Frantar, D. Alistarh
QMoE: Sub-1-Bit Compression of Trillion Parameter Models
Proceedings of Machine Learning and Systems, 2024

[Huang et. al. 25] W. Huang, Y. Liao, J. Liu, R. He, H. Tan, S. Zhang, H. Li, S. Liu, X. QI
Mixture Compressor for Mixture-of-Experts LLMs Gains More
The Thirteenth International Conference on Learning Representations, 2025

References IV

[Li et. al. 24] P. Li, X. Jin, Y. Cheng, T. Chen
Examining post-training quantization for mixture-of-experts: A benchmark
arXiv preprint arXiv:2406.08155, 2024

[Jiang et. al. 24] A. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. Chaplot, D. Casas, E. Hanna, F. Bressand, and others
Mixtral of experts
arXiv preprint arXiv:2401.04088, 2024

[Dong et. al. 24] Z. Dong, Z. Yao, D. Arfeen, A. Gholami, M. W. Mahoney, K. Keutzer.
Hawq-v2: Hessian aware trace-weighted quantization of neural networks
Advances in neural information processing systems, 2020

References V

[Huang et. al. 25] W. Huang, H. Qin, Y. Liu, Y. Li, Q. Liu, X. Liu, L. Benini, M. Magno, S. Zhang, X. Qi.
SliM-LLM: Salience-Driven Mixed-Precision Quantization for Large Language Models
Forty-second International Conference on Machine Learning, 2025