One-Shot Training for Reduction of Performance Variance in LLM Rotation Quantization

Sangki Park
Department of Electronic Engineering
Hanyang University
Seoul, 04763 Korea
skpark1101@hanyang.ac.kr

Chanhoon Kim
Department of Electronic Engineering
Hanyang University
Seoul, 04763 Korea
kch1103@hanyang.ac.kr

Ki-Seok Chung
Department of Electronic Engineering
Hanyang University
Seoul, 04763 Korea
kchung@hanyang.ac.kr

Abstract—**Rotation-based quantization techniques have emerged as effective methods for compressing Large Language Models by mitigating the impact of activation outliers. However, existing approaches face significant challenges: random rotation methods like QuaRot exhibit substantial performance variability across different random seeds, while learned rotation methods like SpinQuant require hundreds of calibration samples, increasing the training cost. To address these limitations, we propose OneRot, a novel one-shot rotation-based quantization technique that achieves stable performance with only a single calibration sample. OneRot introduces random perturbations to layer activations during training, which amplify gradients and enable effective rotation matrix learning in a single attempt while simulating diverse activation distributions. Experimental results on Llama-2 7B, Llama-3.2 3B, and Llama-3.2 1B models with 4-bit quantization demonstrate that OneRot reduces perplexity by 0.25, 2.88, and 21.48, respectively, compared to QuaRot, and reduces standard deviation by 0.01, 0.24, and 1.41, respectively, while achieving comparable performance to SpinQuant using 800 times fewer samples. OneRot provides a practical solution for efficient LLM deployment by combining the stability of learned rotations with the computational efficiency of random approaches.**

Index Terms—**Deep learning, Model compression, LLM, Quantization, Rotation-based quantization.**

I. Introduction

Recently, with the emergence of diverse applications utilizing Large Language Models (LLMs), the efficient deployment and operation of LLMs has become a crucial research topic. In particular, LLMs contain a vast number of parameters, resulting in extremely large model sizes and computational requirements. Among various approaches to address this challenge, quantization techniques have gained significant attention, and numerous studies have demonstrated that they can substantially reduce model size and computational cost while minimizing performance degradation of LLMs.

Among these approaches, rotation-based quantization, which applies rotation transformations to model weights and input activations to reduce quantization error, has recently emerged as an important technique for LLM compression and has shown outstanding performance [1]–[3]. Through rotation, the distribution of tensors can be improved to minimize the impact of outliers that adversely affect quantization distributions, effectively reducing quantization error.

However, existing rotation-based quantization methods have the following limitations: 1) QuaRot [1], a random rotation technique, exhibits significant performance variability even with identical models and hyperparameters due to its inherent randomness. This performance variability can lead to unpredictable results in practical applications, hindering stable model deployment. 2) SpinQuant [2], which learns rotation matrices, requires hundreds of calibration samples for training, resulting in additional training costs.

Therefore, in this paper, we propose OneRot, a novel one-shot training method that overcomes the two aforementioned limitations of rotation-based quantization. OneRot adds random perturbations to input activations during training to make rotation matrix learning effective even with a single training sample. These perturbations amplify gradients, thereby minimizing performance variability across different hyperparameter settings and ensuring consistent performance of quantized models. Furthermore, since OneRot uses only a single sample, it can significantly reduce training cost compared to SpinQuant. To validate the effectiveness of the proposed method, we conducted experiments on various LLMs, and the results confirm that the proposed one-shot training method significantly reduces performance variability in rotation-based quantization and improves overall model performance. The main contributions of OneRot are summarized as follows:

- OneRot proposes a novel one-shot training method that learns rotation matrices using only a single training sample.
- OneRot adds random perturbations to input activations during training, making rotation matrix learning effective even with a single sample.
- Through experiments on various LLMs, we confirm that OneRot significantly reduces performance variability in rotation-based quantization and improves the overall model performance.

II. Background and Related Works

Activation quantization determines quantization parameters using the distribution of a calibration dataset, where one of the critical factors affecting quantization performance is

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00409492)

the presence of outliers. Outliers refer to activation values with significantly larger magnitudes compared to other values, which can distort the quantization range, increase quantization error, and lead to model performance degradation. In particular, for Large Language Models (LLMs), outliers frequently occur in layer activations, negatively impacting quantization performance.

To address this issue, several studies have proposed methods to effectively handle outliers, among which rotation-based quantization techniques have gained considerable attention. Rotation-based quantization methods apply specific rotation transformations to activations or weights to reduce the impact of outliers and minimize quantization error. These rotation-based quantization techniques have the advantage of computational invariance due to the properties of Hadamard matrices or orthogonal matrices, maintaining identical model outputs before and after rotation transformations.

Given an orthogonal matrix Q, activation X, and weight W, the output after rotation transformation is computed as follows:

$$Y = (XQ^T)(QW) = X \cdot Q^T Q \cdot W = XW. \tag{1}$$

As shown in (1), rotation transformations can stabilize the distributions of X and W for quantization while preserving the model's output unchanged.

QuaRot [1] applies rotations using random Hadamard matrices, while SpinQuant [2] learns rotation matrices while maintaining orthogonality to improve quantization performance. Recently, DuQuant [3] proposed a novel approach utilizing dual transformations to disperse outliers.

However, QuaRot [1] exhibits significant performance variability depending on which random Hadamard matrix is selected, even under identical settings [2]. Moreover, SpinQuant [2] requires multiple samples during the rotation matrix learning process, which increases training cost. Therefore, there is a need for methods that can reduce performance variability due to randomness while performing efficient training with fewer samples.

In this paper, we propose OneRot, a novel technique that addresses these challenges by adding random perturbations to layer activations to enhance training robustness and performing one-shot training using only a single sample.

III. Proposed Method: OneRot

In this section, we propose OneRot, a novel technique that reduces performance variability in rotation-based quantization of LLMs by adding random perturbations to layer activations to enhance training robustness and performing one-shot training using only a single sample.

The learning of rotation matrix R is conducted using the Cayley SGD [4] technique, where G represents the model's gradient:

$$R := \left(I - \frac{\eta}{2}Y\right)^{-1}\left(I + \frac{\eta}{2}Y\right)R \tag{2}$$

where η is a hyperparameter and $Y = GR^T - RG^T$. Therefore, the update of R is significantly influenced by the gradient G. We add a random perturbation Δ to activation X to intentionally increase the magnitude of gradient G, enabling substantial updates to R in a single iteration. The perturbed activation $\tilde{X}$ is defined as follows:

$$\tilde{X} = X + \Delta, \quad \Delta = \max(|X|). \tag{3}$$

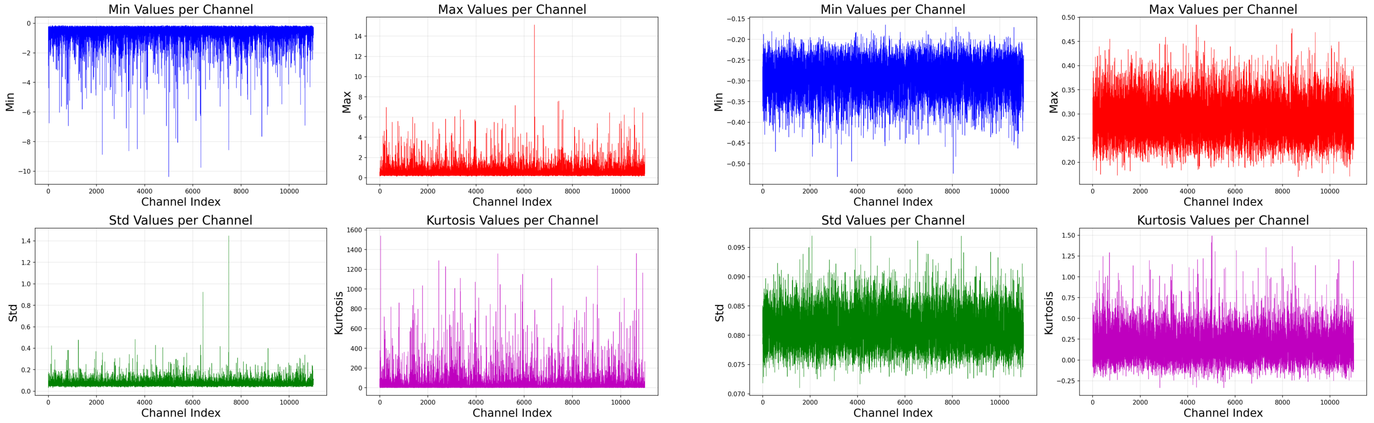
Δ is set to the maximum absolute value of X and is added to n randomly selected elements $x_i \in X$. This perturbation acts similarly to outliers in activations, significantly altering the activation distribution and thereby increasing the gradient G during training, which facilitates substantial updates to the rotation matrix R. Furthermore, the perturbation helps the rotation matrix learn robustly across diverse activation distributions. This achieves effects similar to existing methods that use multiple samples, enabling effective training with a minimal number of samples.

Existing methods, such as SpinQuant [2], require multiple samples, which increases training cost. In contrast, OneRot applies perturbations to a single sample to simulate diverse activation distributions, enabling effective training with only one sample. Through this approach, OneRot dramatically reduces training cost compared to SpinQuant [2], which uses 800 samples, while achieving more stable performance with a rotation overhead comparable to QuaRot [1], which uses random Hadamard matrices.

IV. Experiments

A. Experiment Setup

We conducted experiments on various LLM models and datasets to evaluate the effectiveness of the proposed one-shot training method, OneRot. The models used in the experiments were Llama-2 7B [5], Llama-3.2 3B, and Llama-3.2 1B [6], each compared against QuaRot [1] and SpinQuant [2]. For evaluation, we measured perplexity on WikiText-2 [7] and employed eight zero-shot commonsense reasoning tasks: PIQA [8], HellaSwag [9], WinoGrande [10], ARC-easy and ARC-challenge [11], OpenBookQA [12], BoolQ [13], and SIQA [14]. The learning of rotation matrices was conducted using the Cayley SGD [4] technique. In OneRot, perturbations were added to all layer activations except the first and the last layers, with training performed using a single sample. Furthermore, each experiment was repeated 10 times with different seeds in the same environment, recording the mean and standard deviation. To compare the efficiency of OneRot, we compared it against QuaRot, which uses random Hadamard matrices, and SpinQuant. Each model applied W-A-KV (weight-activation-KV cache) 4-bit quantization, with weights quantized using GPTQ [15] with 128 samples. Following the SpinQuant settings, quantization was performed per-channel for weights (W) and per-token for activations (A), while KV cache used group sizes of 128 for Llama-2 7B and Llama-3.2 3B models, and 64 for the Llama-3.2 1B model. All experiments were implemented using PyTorch and the HuggingFace Transformers library, and executed on NVIDIA RTX 4090 GPUs.



(a) Distributions before applying rotation

(b) Distributions after applying rotation

Fig. 1. Channel-wise distributions of layer activation before and after applying OneRot in Llama-2 7B's layer 15 down projection. (a) shows the distribution of minimum, maximum values, standard deviation, and kurtosis before rotation. (b) shows the same statistics after rotation. This illustrates how OneRot effectively redistributes activation values, reducing outliers and stabilizing the distribution for better quantization performance with only one sample.

TABLE I
RESULT COMPARISON OF WIKITEXT-2 AND ZERO-SHOT COMMONSENSE REASONING TASKS ON LLAMA MODELS WITH W-A-KV 4-4-4 QUANTIZATION. EACH RESULT IS PRESENTED WITH MEAN AND STANDARD DEVIATION.

Model	**Method**	**Dataset**	
		Wiki2(↓)	**0-shot Avg(↑)**
Llama-2 7B	QuaRot	6.37 ± 0.02	60.01 ± 0.39
	SpinQuant	6.12 ± 0.01	60.67 ± 0.32
	Ours	6.12 ± **0.01**	60.71 ± **0.24**
Llama-3.2 3B	QuaRot	12.83 ± 0.26	55.31 ± 0.44
	SpinQuant	9.92 ± 0.03	56.15 ± 0.41
	Ours	9.95 ± **0.02**	56.02 ± **0.37**
Llama-3.2 1B	QuaRot	35.92 ± 1.47	43.06 ± 0.52
	SpinQuant	14.43 ± 0.07	48.20 ± **0.40**
	Ours	14.44 ± **0.06**	48.06 ± 0.58

TABLE II
STANDARD DEVIATION OF ZERO-SHOT COMMONSENSE REASONING TASKS ON LLAMA-2 7B WITH 4-4-4 QUANTIZATION ACCORDING TO THE NUMBER OF CALIBRATION SAMPLES.

Method	**#Sample**	σ
Ours	800	0.22
	1	0.24

B. Experiment Results

Table I presents a comparison of the performance of the proposed OneRot method against existing rotation-based quantization techniques, QuaRot and SpinQuant. Model performance was evaluated through perplexity on the WikiText-2 dataset and average accuracy on zero-shot commonsense reasoning tasks. For Llama-2 7B, Llama-3.2 3B, and Llama-3.2 1B models, OneRot achieved WikiText-2 perplexity reductions of 0.25, 2.88, and 21.48, respectively, compared to QuaRot, with standard deviations also decreasing by 0.01, 0.24, and 1.41, demonstrating that performance improvement and stabilization effects become more pronounced as model size decreases. On zero-shot commonsense reasoning tasks, OneRot achieved average accuracy improvements of 0.70%, 0.71%, and 5.00%, respectively, over QuaRot, with standard deviations decreasing by 0.15%, 0.07%, and -0.06%. Notably, while OneRot achieved a 5 percentage point performance improvement over QuaRot on the Llama-3.2 1B model, the standard deviation increased by 0.06 percentage points. However, examining the task-specific performance in Table III reveals that OneRot achieved more stable performance than QuaRot on most tasks for the Llama-3.2 1B model. Furthermore, when compared to SpinQuant, OneRot showed negligible differences in WikiText-2 perplexity and only slight differences in average accuracy on zero-shot commonsense reasoning tasks, while achieving more stable performance with lower standard deviations in most cases.

Table II shows a comparison of standard deviation with respect to the number of samples for OneRot on the Llama-2 7B model with W-A-KV 4-4-4 quantization. These results demonstrate that even when the number of training samples is reduced to one, the learning of rotation matrices via Cayley SGD is not significantly affected, confirming that OneRot can maintain stable performance with a minimal number of samples. Additionally, OneRot consistently achieved lower standard deviations than SpinQuant regardless of the sample count, demonstrating that it provides more stable performance even with fewer samples.

Fig. 1 visualizes the channel-wise changes in activation distribution before and after applying OneRot in the layer 15 down projection layer of the Llama-2 7B model. Fig. 1(a) shows the activation distribution before rotation, where extreme minimum and maximum values can be observed in certain channels, and most Kurtosis values are also high. In contrast, Fig. 1(b) shows the activation distribution after applying OneRot. Here, the minimum and maximum values are more evenly distributed, and the standard deviation and

TABLE III

Performance comparison of different quantization methods across Zero-shot Commonsense Reasoning tasks on various Llama models. All results are reported with mean and standard deviation over 10 runs with different random seeds.

Model	**Method**	**Dataset**							
		arc_challenge	arc_easy	bool_q	hellaswag	openbookqa	piqa	social_iqa	winogrande
2 7B	QuaRot	40.97±0.73	67.96±1.51	73.80±1.20	71.66±0.39	40.46±0.98	75.87±0.57	43.81±0.37	65.59±1.17
	SpinQuant	42.14±0.81	70.02±0.74	73.70±0.59	72.72±0.23	40.61±0.79	76.15±0.51	44.23±0.68	65.81±1.00
	Ours	41.67±0.79	70.37±0.55	73.95±0.64	72.68±0.21	41.00±1.21	76.18±0.50	44.04±0.70	65.76±0.82
3.2 3B	QuaRot	37.44±1.21	63.16±1.20	62.94±3.10	65.46±0.42	37.92±1.29	72.45±0.50	42.33±0.82	60.71±1.29
	SpinQuant	39.36±0.78	61.81±0.82	63.20±1.69	67.24±0.25	37.80±0.91	73.38±0.61	43.29±0.60	63.14±0.85
	Ours	38.77±1.23	62.68±1.27	62.72±1.65	67.16±0.34	37.84±1.28	73.29±0.55	43.28±0.48	62.44±0.57
3.2 1B	QuaRot	26.73±1.13	45.37±1.16	54.11±2.42	41.52±0.47	27.22±0.85	61.56±1.35	36.17±0.73	51.83±1.57
	SpinQuant	31.06±0.77	51.12±0.77	55.83±1.86	54.30±0.23	32.60±0.82	66.77±0.58	39.25±0.51	54.62±1.14
	Ours	31.31±1.00	51.18±0.94	55.78±2.35	54.12±0.29	31.40±1.12	66.55±0.55	39.49±0.83	54.70±1.27

Kurtosis are significantly reduced, demonstrating that the overall distribution has been stabilized. These changes demonstrate that OneRot can effectively redistribute activation values with only a single sample, reducing outliers and improving quantization performance.

V. Conclusion

In this paper, we propose OneRot, a novel one-shot rotation-based quantization technique that reduces the performance variability and the training cost inherent in existing rotation-based quantization methods for LLMs. By adding random perturbations to layer activations, OneRot makes rotation matrix learning effective even with only a single calibration sample, dramatically reducing the training cost while maintaining stable quantization performance.

Our experimental results demonstrate that OneRot achieves significant improvements over QuaRot across Llama models of various sizes with W-A-KV 4-4-4 quantization. On Llama-2 7B, Llama-3.2 3B, and Llama-3.2 1B models, OneRot reduced WikiText-2 perplexity by 0.25, 2.88, and 21.48, respectively, with corresponding reductions in standard deviation, demonstrating enhanced stability, particularly for smaller models. Additionally, OneRot achieved average accuracy improvements of 0.70%, 0.71%, and 5.00% on zero-shot commonsense reasoning tasks compared to QuaRot, while maintaining comparable performance to SpinQuant with substantially lower training cost—using only one sample instead of 800.

OneRot provides a practical and efficient solution for deploying quantized LLMs by achieving the stability of learned rotation methods while maintaining the computational efficiency of random rotation approaches. Future work will explore extending OneRot to other quantization bit-widths and investigating adaptive perturbation strategies for better performance across different model architectures and layer characteristics.

References

[1] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman, "Quarot: Outlier-free 4-bit inference in rotated llms," in *Neural Information Processing Systems*, 2024.

[2] Z. Liu, C. Zhao, I. Fedorov, B. Soran, D. Choudhary, R. Krishnamoorthi, V. Chandra, Y. Tian, and T. Blankevoort, "Spinquant: LLM quantization with learned rotations," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=ogO6DGE6FZ

[3] H. Lin, H. Xu, Y. Wu, J. Cui, Y. Zhang, L. Mou, L. Song, Z. Sun, and Y. Wei, "Duquant: Distributing outliers via dual transformation makes stronger quantized llms," in *Neural Information Processing Systems*, 2024.

[4] J. Li, F. Li, and S. Todorovic, "Efficient riemannian optimization on the stiefel manifold via the cayley transform," in *International Conference on Learning Representations*, 2020.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[6] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[7] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," 2016.

[8] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[9] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[10] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "An adversarial winograd schema challenge at scale," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199370376

[11] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv:1803.05457v1*, 2018.

[12] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *EMNLP*, 2018.

[13] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[14] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, "Socialiqa: Commonsense reasoning about social interactions," 2019. [Online]. Available: https://arxiv.org/abs/1904.09728

[15] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," in *arXiv.org*, 2022.