Novelty Assessment Report

Paper: ParoQuant: Pairwise Rotation Quantization for Efficient Reasoning LLM Inference
PDF URL: https://openreview.net/pdf?id=1USeVjsKau
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Weight-only post-training quantization (PTQ) compresses the weights of Large Language Models (LLMs) into low-precision representations to reduce memory footprint and accelerate inference. However, the presence of outliers in weights and activations often leads to large quantization errors and severe accuracy degradation, especially in recent reasoning LLMs where errors accumulate across long chains of thought. Existing PTQ methods either fail to sufficiently suppress outliers or introduce significant overhead during inference. In this paper, we propose Pairwise Rotation Quantization (ParoQuant), a weight-only PTQ method that combines hardware-efficient and optimizable independent Givens rotations with channel-wise scaling to even out the magnitude across channels and narrow the dynamic range within each quantization group. We also co-design the inference kernel to fully exploit GPU parallelism and keep the rotations and scaling lightweight at runtime. ParoQuant achieves an average 2.4% accuracy improvement over AWQ on reasoning tasks with less than 10% overhead. This paves the way for more efficient and accurate deployment of reasoning LLMs.

Core Task Landscape

This paper addresses: **Weight-Only Post-Training Quantization for Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Outlier-Aware and Salience-Based Quantization Methods**
- **Distribution Transformation and Smoothing Techniques**
- **Optimization-Based Quantization Parameter Learning**
- **Extreme Low-Bit Quantization Methods**
- **Integer-Only and Hardware-Efficient Quantization**
- **Mixed-Precision and Adaptive Quantization Strategies**
- **Quantization-Aware Training and Fine-Tuning Extensions**
- **Unified Frameworks and Compression Integration**
- **Empirical Analysis and Benchmarking Studies**

Complete Taxonomy Tree

- Weight-Only Post-Training Quantization for Large Language Models Survey Taxonomy
- Outlier-Aware and Salience-Based Quantization Methods
  - Gradient and Hessian-Based Salience Detection (2 papers)
  - [1] Gwq: Gradient-aware weight quantization for large language models (Shao Yihua, 2024) View paper
  - [11] Aptq: Attention-aware post-training mixed-precision quantization for large language models (Ziyi Guan, 2024) View paper
  - Activation-Aware Outlier Handling (3 papers)
  - [3] The super weight in large language models (Yu Mengxia, 2024) View paper
  - [4] Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models (Changhun Lee, 2024) View paper
  - [5] Awq: Activation-aware weight quantization for on-device llm compression and acceleration (Ji Lin, 2024) View paper
  - Structured and Unstructured Outlier Masking (2 papers)
  - [28] PTQ1.61: Push the Real Limit of Extremely Low-Bit Post-Training Quantization Methods for Large Language Models (Zhao Jiaqi, 2025) View paper
  - [49] Layer-Wise High-Impact Parameter Ratio Optimization in Post-Training Quantization for Large Language Models (Cuong Pham, 2025) View paper
- Distribution Transformation and Smoothing Techniques
  - Activation Smoothing and Migration (2 papers)
  - [9] Smoothquant: Accurate and efficient post-training quantization for large language models (Xiao, 2023) View paper
  - [12] Qllm: Accurate and efficient low-bitwidth quantization for large language models (Liu Jing, 2023) View paper
  - Channel Reordering and Clustering (2 papers)
  - [10] Rptq: Reorder-based post-training quantization for large language models (Yuan, 2023) View paper
  - [25] CrossQuant: A Post-Training Quantization Method with Smaller Quantization Kernel for Precise Large Language Model Compression (Liu Wenyuan, 2024) View paper
  - Rotation-Based Distribution Equalization ★ (2 papers)
  - [0] ParoQuant: Pairwise Rotation Quantization for Efficient Reasoning LLM Inference (Anon et al., 2026) View paper
  - [44] DAQ: Density-Aware Post-Training Weight-Only Quantization For LLMs (Luo, 2024) View paper

- Optimization-Based Quantization Parameter Learning
  - Block-Wise Reconstruction Optimization (2 papers)
  - [2] Omniquant: Omnidirectionally calibrated quantization for large language models (Shao, 2023) View paper
  - [14] Septq: A simple and effective post-training quantization paradigm for large language models (Han Liu, 2025) View paper
  - Low-Rank and Decomposition-Based Optimization (3 papers)
  - [26] Dl-qat: Weight-decomposed low-rank quantization-aware training for large language models (Wenjing Ke, 2024) View paper
  - [39] LRQuant: A Unified and Learnable Framework to Post-training Quantization for Transformer-based Large Foundation Models (Chao Zeng, 2025) View paper
  - [41] LRQ: Optimizing Post-Training Quantization for Large Language Models by Learning Low-Rank Weight-Scaling Matrices (Lee Jung Hyun, 2024) View paper
  - Learnable Scaling and Dynamic Range Adjustment (2 papers)
  - [38] Norm Tweaking: High-Performance Low-Bit Quantization of Large Language Models (Li Liang, 2024) View paper
  - [48] Enhancing computation efficiency in large language models through weight and activation quantization (Lee, 2023) View paper
- Extreme Low-Bit Quantization Methods
  - Binary Quantization Approaches (4 papers)
  - [7] Onebit: Towards extremely low-bit large language models (Xu, 2024) View paper
  - [8] Billm: Pushing the limit of post-training quantization for llms (Huang Wei, 2024) View paper
  - [24] PT-BitNet: Scaling up the 1-Bit large language model with post-training quantization (Yufei Guo, 2025) View paper
  - [33] ARB-LLM: Alternating Refined Binarizations for Large Language Models (Li Zhiteng, 2024) View paper
  - Ternary and Sub-2-Bit Quantization (2 papers)
  - [16] PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models (Xiao He, 2025) View paper
  - [36] Rethinking Output Alignment For 1-bit Post-Training Quantization of Large Language Models (Dung Anh Hoang, 2025) View paper
- Integer-Only and Hardware-Efficient Quantization
  - Integer-Only Inference Optimization (1 papers)
  - [13] I-llm: Efficient integer-only inference for fully-quantized low-bit large language models (Hu Xing, 2024) View paper
  - Specialized Format Quantization (4 papers)
  - [19] Post training quantization of large language models with microscaling formats (Sharify, 2024) View paper
  - [22] PoTPTQ: A Two-step Power-of-Two Post-training for LLMs (Wang, 2025) View paper
  - [30] FP4-Quantization: Lossless 4bit Quantization for Large Language Models (Jie Wang, 2024) View paper
  - [37] F-BFQ: Flexible Block Floating-Point Quantization Accelerator for LLMs (Haris, 2025) View paper
- Mixed-Precision and Adaptive Quantization Strategies
  - Sensitivity-Based Mixed-Precision Assignment (1 papers)
  - [17] FBQuant: FeedBack Quantization for Large Language Models (Liu, 2025) View paper
  - Automated Mixed-Precision Search (1 papers)
  - [46] AMQ: Enabling AutoML for Mixed-precision Weight-Only Quantization of Large Language Models (Lee SangJun, 2025) View paper
- Quantization-Aware Training and Fine-Tuning Extensions (1 papers)
  - [6] Llm-qat: Data-free quantization aware training for large language models (Liu, 2024) View paper
- Unified Frameworks and Compression Integration (3 papers)
  - [31] Improving quantization with post-training model expansion (Franco Giuseppe, 2025) View paper
  - [35] SLiM: One-shot Quantization and Sparsity with Low-rank Approximation for LLM Weight Compression (Mozaffari, 2024) View paper
  - [45] Treasures in Discarded Weights for LLM Quantization (Hao Yu, 2025) View paper
- Empirical Analysis and Benchmarking Studies
  - Comparative Method Evaluation (4 papers)
  - [15] Evaluating quantized large language models (Li Shiyao, 2024) View paper
  - [20] Benchmarking post-training quantization in llms: Comprehensive taxonomy, unified evaluation, and comparative analysis (Zhao Jiaqi, 2025) View paper
  - [21] A Comprehensive Study on Post-Training Quantization for Large Language Models (Yao, 2023) View paper
  - [32] Comparison of Weight-Only Quantization (WOQ) models in Large Language Models (Tiwari, 2025) View paper
  - Scaling Laws and Predictive Analysis (2 papers)
  - [34] Scaling Laws for Post Training Quantized Large Language Models (Xu Zifei, 2024) View paper
  - [50] Scaling Laws for Task-Stratified Knowledge in Post-Training Quantized Large Language Models (Zhou Chen-xi, 2025) View paper
  - Task-Specific and Trade-Off Analysis (7 papers)
  - [23] Exploring the Trade-Offs: Quantization Methods, Task Difficulty, and Model Size in Large Language Models From Edge to Giant (Jemin Lee, 2024) View paper
  - [27] Understanding the impact of post-training quantization on large language models (Roy Somnath, 2023) View paper
  - [29] Interactions Across Blocks in Post-Training Quantization of Large Language Models (Khasmamad Shabanovi, 2024) View paper
  - [40] Optimizing Large Language Models through Quantization: A Comparative Analysis of PTQ and QAT Techniques (Hasan, 2024) View paper
  - [42] â□□Give Me BF16 or Give Me Deathâ□□? Accuracy-Performance Trade-Offs in LLM Quantization (Kurtic, 2025) View paper
  - [43] Understanding the difficulty of low-precision post-training quantization of large language models (Xu Zifei, 2024) View paper
  - [47] Domain aware post training quantization for vision transformers in deployment (Li Wang, 2025) View paper
  - Survey and Taxonomy Papers (1 papers)
  - [18] A survey on 1-bit quantized large language models (Kritika Tripathi, 2025) View paper

Narrative

Core task: weight-only post-training quantization for large language models. The field has organized itself around several complementary strategies for compressing LLM weights without retraining. Outlier-aware and salience-based methods such as AWQ[5] and OWQ[4] identify and protect critical weights that disproportionately affect model accuracy. Distribution transformation techniques like SmoothQuant[9] and rotation-based approaches including RPTQ[10] reshape weight distributions to make them more amenable to low-bit representation. Optimization-based methods learn quantization parameters through careful calibration, while extreme low-bit

quantization pushes toward ternary or binary weights as seen in BiLLM[8] and OneBit[7]. Integer-only and hardware-efficient designs target deployment constraints, mixed-precision strategies allocate bits adaptively across layers, and unified frameworks integrate quantization with other compression techniques. Empirical studies such as Benchmarking PTQ[20] and Evaluating Quantized LLMs[15] systematically compare these diverse approaches.

A central tension runs between methods that rely on careful weight selection versus those that transform the entire distribution. Rotation-based equalization methods like RPTQ[10] and DAQ[44] apply learned or fixed rotations to homogenize weight magnitudes before quantization, reducing the burden on per-channel scaling. ParoQuant[0] sits squarely in this rotation-based cluster, emphasizing distribution equalization through orthogonal transformations. Compared to salience-driven approaches like AWQ[5] that preserve outliers through non-uniform scaling, ParoQuant[0] and its rotation-based neighbors pursue a more global reshaping strategy. This contrasts with optimization-heavy methods such as OmniQuant[2] that jointly tune multiple quantization parameters, and with extreme quantization works like BiLLM[8] that accept higher approximation error in exchange for maximal compression. The rotation paradigm offers a middle ground: it avoids expensive per-weight decisions while achieving better uniformity than naive round-to-nearest schemes.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. DAQ: Density-Aware Post-Training Weight-Only Quantization For LLMs

Authors: Luo, Yingsong, Chen Ling, YINGâ□□JIA Luo, Ling Chen | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large language models (LLMs) excel in various tasks but face deployment challenges due to hardware constraints. We propose density-aware post-training weight-only quantization (DAQ), which has two stages: 1) density-centric alignment, which identifies the center of high-density weights and centers the dynamic range on this point to align high-density weight regions with floating-point high-precision regions; 2) learnable dynamic range adjustment, which adjusts the dynamic range by optimizing qua...

Relationship Analysis

Both papers belong to the Rotation-Based Distribution Equalization category, applying transformations to equalize weight magnitudes before quantization. They overlap in using rotation-based approaches to address outlier issues in weight-only quantization for LLMs. However, ParoQuant focuses on hardware-efficient independent Givens rotations combined with channel-wise scaling, optimizing for minimal inference overhead through algorithm-system co-design, while DAQ emphasizes density-aware alignment that centers dynamic ranges on high-density weight regions followed by learnable dynamic range adjustment, without employing rotation transformations.

Contributions Analysis

Overall novelty summary. ParoQuant proposes a weight-only post-training quantization method combining independent Givens rotations with channel-wise scaling to equalize weight distributions before quantization. The paper resides in the 'Rotation-Based Distribution Equalization' leaf, which contains only two papers including ParoQuant itself and one sibling (RPTQ). This represents a relatively sparse research direction within the broader taxonomy of fifty papers across thirty-six topics, suggesting rotation-based approaches constitute a focused but less crowded area compared to outlier-aware methods or optimization-based techniques.

The rotation-based leaf sits within the 'Distribution Transformation and Smoothing Techniques' branch, which also includes activation smoothing methods like SmoothQuant and channel reordering approaches. Neighboring branches pursue different strategies: outlier-aware methods explicitly preserve salient weights through mixed-precision or masking, while optimization-based approaches learn quantization parameters through reconstruction objectives. ParoQuant's rotation paradigm diverges from these by globally reshaping distributions rather than selectively protecting outliers or iteratively optimizing parameters, positioning it as a complementary approach to salience-driven techniques like AWQ.

Among twenty-eight candidates examined through semantic search and citation expansion, the core ParoQuant method shows overlap with two prior works, while the scaled pairwise rotation transform and co-designed inference kernel examined ten and nine candidates respectively with no clear refutations. The method-level analysis suggests that while rotation-based quantization has precedent in the limited search scope, the specific combination of pairwise Givens rotations with hardware-efficient kernel design may represent a less-explored configuration. The statistics indicate moderate prior work density for the overall approach but sparser coverage for the implementation-focused contributions.

Based on the limited search scope of top-thirty semantic matches, ParoQuant appears to occupy a moderately novel position within rotation-based quantization, though the small size of this research direction makes comprehensive novelty assessment challenging. The analysis covers rotation-based and distribution transformation methods but may not capture all relevant work in hardware-efficient quantization or kernel optimization, which could provide additional context for the inference kernel contribution.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Pairwise Rotation Quantization (ParoQuant) method

Description: ParoQuant is a weight-only post-training quantization method that uses independent Givens rotations combined with channel-wise scaling to suppress outliers in weights. This transform evens out magnitude across channels and narrows the dynamic range within quantization groups, making weights more quantization-friendly.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. BASE-Q: Bias and Asymmetric Scaling Enhanced Rotational Quantization for Large Language Models

URL: View paper

Brief Assessment

BASE-Q[59] focuses on addressing limitations of existing rotational quantization methods through bias correction and asymmetric scaling, rather than proposing a new rotation-based quantization method like ParoQuant.

---

2. Rolora: Fine-tuning rotated outlier-free llms for effective weight-activation quantization

URL: View paper

Brief Assessment

RoLora[68] focuses on integrating rotation with LoRA fine-tuning for weight-activation quantization, while ParoQuant is a weight-only post-training quantization method. The technical approaches differ fundamentally in their application context (fine-tuning vs. post-training only).

---

3. Duquant: Distributing outliers via dual transformation makes stronger quantized llms

URL: View paper

Prior Art Analysis

Duquant[62] demonstrates that rotation-based transformations for post-training quantization with outlier suppression were already explored before ParoQuant. Both papers use rotation matrices combined with channel-wise scaling to redistribute outliers and make weights more quantization-friendly. Duquant[62] employs block-diagonal rotation matrices constructed using greedy search with prior knowledge of outlier dimensions, followed by permutation transformations, to handle both normal and massive outliers. The paper explicitly discusses rotation transformations as an established technique and compares against prior rotation-based methods, indicating that the core concept of using rotations with scaling for outlier suppression in quantization was not novel to ParoQuant.

Evidence

Evidence 1 - **Rationale**: Both papers propose rotation-based transformations combined with additional techniques (ParoQuant uses channel-wise scaling, Duquant[62] uses permutation) to redistribute outliers for quantization. This shows the rotation approach for outlier suppression was already established. - **Original**: we proposepairwise rotation quantization (paroquant), a weight-only ptq method that combines hardware-efficient and optimizable independent givens rotations with channelwise scaling to even out the magnitude across channels and narrow the dynamic range within each quantization group. - **Candidate**: we propose the duquant method, which includes the rotation and permutation transformations based on the smooth technique. By combining rotation transformation and channel permutation, our duquant method aims to redistribute these features within the activation space, thereby mitigating the effects o...

Evidence 2 - **Rationale**: Both papers use rotation matrices to narrow dynamic range and suppress outliers. Duquant[62] explicitly constructs rotation matrices targeting outlier dimensions, demonstrating prior work on rotation-based outlier suppression. - **Original**: channel-wise scaling evens out the average magnitude across channels, while the pairwise (i.e., givens) rotations align the values within each channel pair at every token position, narrowing the dynamic range of each quantization group. - **Candidate**: To address this problem, we employ a greedy search with prior knowledge to compute a rotation matrix ˆr, thereby approximating the ideal rotation matrix r. Specifically, the calculation of ˆr involves the following steps, ◦ identify the feature dimension d(1) where the outlier are primarily concentr...

Evidence 3 - **Rationale**: The original paper acknowledges rotation transforms as existing prior work. Duquant[62] uses block-diagonal rotation matrices for outlier redistribution, confirming that rotation-based approaches for quantization were already established in the field. - **Original**: two main types of transform have been proposed in previous work: channel-wise scaling, where t is a diagonal matrix (lin et al., 2024b; shao et al., 2024; wei et al., 2023), and rotation, where t is an orthogonal matrix (chee et al., 2023; ashkboos et al., 2024; liu et al., 2025b; lin et al., 2024a;... - **Candidate**: To effectively mitigate both normal and massive outliers, we first employ the smooth technique to shift the quantization challenge from activations to weights. Next, we introduce the block-diagonal rotation matrix ˆr to locally redistribute feature outliers within the activation space.

Evidence 4 - **Rationale**: Both papers discuss applying rotation transformations to smooth weights and activations for quantization. This demonstrates that the concept of using rotations with channel-wise operations for outlier suppression was already explored before ParoQuant. - **Original**: rotation enables crosschannel interactions that can concentrate values more effectively than channel-wise scaling (chee et al., 2023; liu et al., 2025b). however, rotations cannot be merged into element-wise operators (e.g., layer normalization) like channel-wise scaling does, so they usually requir... - **Candidate**: remark 1. it is worth noting that the proposed duquant method can simultaneously smooth the weight matrix. While the commonly adopted smooth technique is effective, it can cause the weight matrix of the down-projection layer to exhibit pronounced outliers, leading to performance degradation. however...

---

4. Quarot: Outlier-free 4-bit inference in rotated llms

URL: View paper

Prior Art Analysis

Quarot[61] demonstrates that rotation-based transforms for outlier suppression in quantization were published prior to the ORIGINAL paper's submission. Quarot[61] uses randomized Hadamard transformations (a type of rotation matrix) combined with computational invariance to eliminate outliers in weights and activations for 4-bit quantization. The paper explicitly applies orthogonal rotations to weight matrices and shows that this removes outliers, making quantization easier. This establishes that the core concept of using rotation transforms with scaling for outlier suppression in post-training quantization was already known and published before the ORIGINAL paper.

Evidence

Evidence 1 - **Rationale**: Both papers use rotation transforms to address outlier features in quantization. Quarot[61] uses Hadamard rotations (orthogonal matrices) while the original uses Givens rotations, but both achieve the same goal of outlier suppression through rotation. - **Original**: channel-wise scaling evens out the average magnitude across channels, while the pairwise (i.e., givens) rotations align the values within each channel pair at every token position, narrowing the dynamic range of each quantization group. - **Candidate**: we address the issue of outlier features by rotating the inputs of the model using randomized hadamard transformations. we do this using the computational invariance idea [ashkboos et al. , 2024] and fuse hadamard transformations into the weight matrices, resulting in an equivalent network without o...

Evidence 2 - **Rationale**: Quarot[61] demonstrates that rotations (Hadamard transformations) eliminate outlier features, establishing prior work on the effectiveness of rotations for outlier suppression in quantization. - **Original**: rotations effectively suppress outliers, and (2) a sparsely parameterized rotation can be as effective as a full rotation. - **Candidate**: this enables the weights, activations, and kv caches to be quantized to 4 bits with minimal accuracy drop. our main contributions are: • we show that randomized hadamard transformations can be applied to the weight matrices without additional model modifications. in turn, this completely eliminates o...

Evidence 3 - **Rationale**: Quarot[61] describes absorbing scaling operations into weight matrices, which is conceptually similar to the channel-wise scaling component of ParoQuant that evens out magnitudes across channels. - **Original**: channel-wise scaling evens out the magnitude across channels and can usually be merged into preceding operators without incurring extra overhead - **Candidate**: if win is a weight matrix that appears on the left of a transformer block (i.e., wgate, wup in figure 2, or wk, wq, wv in figure 5) then we can multiply on the left by an orthogonal matrix q, and cancel out this effect by multiplying the output matrix (wdown, wout) by q⊤. this applies despite the f...

---

5. Rotated Runtime Smooth: Training-Free Activation Smoother for accurate INT4 inference

URL: View paper

Brief Assessment

Rotated Runtime Smooth[67] focuses on runtime activation smoothing with rotation operations applied to activations during inference, not weight-only quantization with optimized Givens rotations applied to weights as in ParoQuant.

---

6. SmoothRot: Combining Channel-Wise Scaling and Rotation for Quantization-Friendly LLMs

URL: View paper

Brief Assessment

SmoothRot[64] focuses on combining channel-wise scaling with Hadamard transformations for activation quantization in FFN modules, while ParoQuant uses optimizable independent Givens rotations for weight-only quantization. The technical approaches and application domains differ fundamentally.

---

7. A Comprehensive Evaluation on Quantization Techniques for Large Language Models

URL: View paper

Brief Assessment

Comprehensive Evaluation Quantization[55] is a survey paper that evaluates existing quantization methods but does not propose ParoQuant. It discusses rotation-based methods generically (e.g., QuIP, QuaRot, SpinQuant) without claiming to introduce the specific ParoQuant method or its independent Givens rotations combined with channel-wise scaling design.

---

8. Turning LLM Activations Quantization-Friendly

URL: View paper

Brief Assessment

Turning Activations Friendly[63] focuses on weight-activation quantization with a hybrid approach combining channel-wise scaling before rotation, while ParoQuant is a weight-only method using independent Givens rotations with channel-wise scaling for outlier suppression in weights.

---

9. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations

URL: View paper

Brief Assessment

RotateKV[65] focuses on KV cache quantization for LLM inference using rotation techniques, while ParoQuant addresses weight-only quantization. The technical domains and application targets are fundamentally different.

Contribution 2: Scaled pairwise rotation transform

Description: The scaled pairwise rotation is a novel transform that applies a series of independent Givens rotations (pairwise rotations with no dependencies) combined with channel-wise scaling. This design enables effective outlier suppression while maintaining computational efficiency through GPU parallelism.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. BTC-LLM: Efficient Sub-1-Bit LLM Quantization via Learnable Transformation and Binary Codebook

URL: View paper

Brief Assessment

BTC-LLM[60] uses learnable scaling and rotation matrices for binary quantization, while the original paper applies Givens rotations with channel-wise scaling for weight-only quantization. These are fundamentally different quantization paradigms (binary vs. low-bit linear) with distinct technical approaches.

---

2. BASE-Q: Bias and Asymmetric Scaling Enhanced Rotational Quantization for Large Language Models

URL: View paper

Brief Assessment

BASE-Q[59] does not propose Givens rotations or pairwise rotation transforms. Instead, it identifies limitations of rotation methods and proposes bias correction and asymmetric scaling as solutions.

---

3. Deep learning image compression with multi-channel tANS coding and hardware deployment

URL: View paper

Brief Assessment

Multi-channel tANS[57] focuses on image compression using convolutional neural networks with KLT-based transforms and sequential encoding. This is fundamentally different from the original paper's scaled pairwise rotation for LLM weight quantization.

---

4. Mixture attention block and Swin transformerâ□□based entropy model for learned image compression

URL: View paper

Brief Assessment

Mixture Attention Block[54] focuses on image compression using attention mechanisms and entropy modeling for spatial-channel feature optimization, not on Givens rotations or channel-wise scaling transforms for weight quantization in LLMs.

---

5. Quantization Methods for Matrix Multiplication and Efficient Transformers

URL: View paper

Brief Assessment

Matrix Multiplication Quantization[58] focuses on nested lattice codebooks for vector quantization of matrix products, not on Givens rotations or channel-wise scaling transforms for weight quantization.

---

6. A Comprehensive Evaluation on Quantization Techniques for Large Language Models

URL: View paper

Brief Assessment

Comprehensive Evaluation Quantization[55] reviews rotation and scaling techniques separately as pre-quantization transformations, but does not describe or claim the specific scaled pairwise rotation transform design with independent Givens rotations. The paper evaluates combinations of existing methods rather than proposing this novel transform architecture.

---

7. Systematic codebook designs for quantized beamforming in correlated MIMO channels

URL: View paper

Brief Assessment

Systematic Codebook Designs[56] focuses on rotation and scaling maps for quantizing beamforming vectors in MIMO channels, not weight quantization for LLMs. The technical domains and applications are fundamentally different.

---

8. Color conversion matrices in digital cameras: a tutorial

URL: View paper

Brief Assessment

Color Conversion Matrices[53] focuses on color rotation matrices for digital camera image processing in color space conversion, not on weight quantization or outlier suppression in neural networks.

---

9. Neural Networks with Model Compression

URL: View paper

Brief Assessment

Model Compression[52] mentions using 'an extra scale factor on the activation' and 'orthogonal binary codes' but provides insufficient detail about the transform architecture. The candidate's sparse context does not demonstrate prior work on combining independent Givens rotations with channel-wise scaling for outlier suppression in weight quantization.

---

10. OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting

URL: View paper

Brief Assessment

OstQuant[51] uses orthogonal and scaling transformations but does not employ Givens rotations or pairwise rotation mechanisms. The candidate focuses on general orthogonal transformations optimized via QSUR metric, while the original paper specifically designs independent Givens rotations for GPU parallelism.

Contribution 3: Co-designed inference kernel for efficient transform computation

Description: The authors developed a specialized CUDA kernel that exploits three levels of parallelism (token, channel group, and pair) to efficiently compute the scaled pairwise rotation transform during inference. This system-level design ensures minimal overhead while maintaining quantization accuracy.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Q-Palette: Fractional-Bit Quantizers Toward Optimal Bit Allocation for Efficient LLM Deployment

URL: View paper

Brief Assessment

Q-Palette[75] focuses on fractional-bit quantizers with rotation-based preprocessing and optimized CUDA kernels for dequantization and matrix multiplication, not on GPU-parallel kernels specifically for rotation-based quantization transforms during inference as described in the original paper's pairwise rotation approach.

---

2. Rotate, Clip, and Partition: Towards W2A4KV4 Quantization by Integrating Rotation and Learnable Non-uniform Quantizer

URL: View paper

Brief Assessment

Rotate Clip Partition[69] focuses on LUT-based GEMV kernels for non-uniform quantization with learnable partitioning, not on parallelized rotation transform kernels. The candidate's kernel design addresses different computational challenges (non-uniform dequantization) than the original's GPU-parallel rotation computation.

---

3. ConvRot: Rotation-Based Plug-and-Play 4-bit Quantization for Diffusion Transformers

URL: View paper

Brief Assessment

ConvRot[71] focuses on diffusion transformers with a plug-and-play module (ConvLinear4bit) that fuses rotation, quantization, GEMM, and dequantization. The original paper targets LLM inference with a specialized CUDA kernel exploiting token, channel group, and pair-level parallelism for Givens rotations. These are distinct system designs for different model architectures and computational patterns.

---

4. Pushing the Limits of Large Language Model Quantization via the Linearity Theorem

URL: View paper

Brief Assessment

Linearity Theorem[76] focuses on GPU kernels for lookup-table-based quantization grids (FLUTE adaptation), not rotation-based transforms. The candidate's kernel design addresses vectorized indexing into quantization grids, while the original paper's kernel exploits parallelism for scaled pairwise rotation transforms during inference.

---

5. ROSAQ: Rotation-based Saliency-Aware Weight Quantization for Efficiently Compressing Large Language Models

URL: View paper

Brief Assessment

ROSAQ[74] focuses on PCA-based rotation with mixed-precision quantization and kernel fusion for speedup, while the original paper develops specialized CUDA kernels for pairwise Givens rotations with three-level parallelism (token, channel group, pair). These are fundamentally different rotation mechanisms and kernel designs.

---

6. TR-DQ: Time-Rotation Diffusion Quantization

URL: View paper

Brief Assessment

TR-DQ[70] focuses on diffusion model quantization with rotation matrices for time-step-aware processing, not on GPU-parallel kernels for rotation-based quantization in LLMs. The technical domains and optimization targets differ fundamentally.

---

7. Bridging the Gap Between Promise and Performance for Microscaling FP4 Quantization

URL: View paper

Brief Assessment

Bridging Gap Microscaling[72] focuses on GPU kernels for microscaling FP4 quantization with Hadamard transforms and block-wise operations, not on pairwise Givens rotations with three-level parallelism (token, channel group, pair) as described in the original paper's CUDA kernel design.

---

8. Breaking the Efficiency-Accuracy: Fusion of Rotation Quantization and N: M Sparsity for LLMs Inference

URL: View paper

Brief Assessment

Breaking Efficiency-Accuracy[77] focuses on fusing rotation quantization with N:M sparsity and implements custom CUDA kernels for sparse operations, not specifically for rotation-based quantization transforms. The kernel design targets different optimization goals (sparsity allocation) compared to the original paper's pairwise rotation transform kernel.

---

9. KVLinC: KV Cache Quantization with Hadamard Rotation and Linear Correction

URL: View paper

Brief Assessment

KVLinC[73] focuses on KV cache quantization with a custom attention kernel for decoding, not on rotation-based weight quantization transforms. The kernel design addresses different computational patterns (attention over quantized cache) rather than the scaled pairwise rotation transform for weight matrices described in the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] ParoQuant: Pairwise Rotation Quantization for Efficient Reasoning LLM Inference View paper
- [1] Gwq: Gradient-aware weight quantization for large language models View paper
- [2] Omniquant: Omnidirectionally calibrated quantization for large language models View paper
- [3] The super weight in large language models View paper
- [4] Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models View paper
- [5] Awq: Activation-aware weight quantization for on-device llm compression and acceleration View paper
- [6] Llm-qat: Data-free quantization aware training for large language models View paper
- [7] Onebit: Towards extremely low-bit large language models View paper
- [8] Billm: Pushing the limit of post-training quantization for llms View paper
- [9] Smoothquant: Accurate and efficient post-training quantization for large language models View paper
- [10] Rptq: Reorder-based post-training quantization for large language models View paper
- [11] Aptq: Attention-aware post-training mixed-precision quantization for large language models View paper
- [12] Qllm: Accurate and efficient low-bitwidth quantization for large language models View paper
- [13] I-llm: Efficient integer-only inference for fully-quantized low-bit large language models View paper
- [14] Septq: A simple and effective post-training quantization paradigm for large language models View paper
- [15] Evaluating quantized large language models View paper
- [16] PTQTP: Post-Training Quantization to Trit-Planes for Large Language Models View paper
- [17] FBQuant: FeedBack Quantization for Large Language Models View paper
- [18] A survey on 1-bit quantized large language models View paper
- [19] Post training quantization of large language models with microscaling formats View paper
- [20] Benchmarking post-training quantization in llms: Comprehensive taxonomy, unified evaluation, and comparative analysis View paper
- [21] A Comprehensive Study on Post-Training Quantization for Large Language Models View paper
- [22] PoTPTQ: A Two-step Power-of-Two Post-training for LLMs View paper
- [23] Exploring the Trade-Offs: Quantization Methods, Task Difficulty, and Model Size in Large Language Models From Edge to Giant View paper
- [24] PT-BitNet: Scaling up the 1-Bit large language model with post-training quantization View paper
- [25] CrossQuant: A Post-Training Quantization Method with Smaller Quantization Kernel for Precise Large Language Model Compression View paper
- [26] Dl-qat: Weight-decomposed low-rank quantization-aware training for large language models View paper
- [27] Understanding the impact of post-training quantization on large language models View paper
- [28] PTQ1.61: Push the Real Limit of Extremely Low-Bit Post-Training Quantization Methods for Large Language Models View paper
- [29] Interactions Across Blocks in Post-Training Quantization of Large Language Models View paper
- [30] FP4-Quantization: Lossless 4bit Quantization for Large Language Models View paper
- [31] Improving quantization with post-training model expansion View paper
- [32] Comparison of Weight-Only Quantization (WOQ) models in Large Language Models View paper
- [33] ARB-LLM: Alternating Refined Binarizations for Large Language Models View paper
- [34] Scaling Laws for Post Training Quantized Large Language Models View paper
- [35] SLiM: One-shot Quantization and Sparsity with Low-rank Approximation for LLM Weight Compression View paper
- [36] Rethinking Output Alignment For 1-bit Post-Training Quantization of Large Language Models View paper
- [37] F-BFQ: Flexible Block Floating-Point Quantization Accelerator for LLMs View paper
- [38] Norm Tweaking: High-Performance Low-Bit Quantization of Large Language Models View paper
- [39] LRQuant: A Unified and Learnable Framework to Post-training Quantization for Transformer-based Large Foundation Models View paper
- [40] Optimizing Large Language Models through Quantization: A Comparative Analysis of PTQ and QAT Techniques View paper
- [41] LRQ: Optimizing Post-Training Quantization for Large Language Models by Learning Low-Rank Weight-Scaling Matrices View paper
- [42] â□□Give Me BF16 or Give Me Deathâ□□? Accuracy-Performance Trade-Offs in LLM Quantization View paper
- [43] Understanding the difficulty of low-precision post-training quantization of large language models View paper
- [44] DAQ: Density-Aware Post-Training Weight-Only Quantization For LLMs View paper
- [45] Treasures in Discarded Weights for LLM Quantization View paper
- [46] AMQ: Enabling AutoML for Mixed-precision Weight-Only Quantization of Large Language Models View paper
- [47] Domain aware post training quantization for vision transformers in deployment View paper
- [48] Enhancing computation efficiency in large language models through weight and activation quantization View paper
- [49] Layer-Wise High-Impact Parameter Ratio Optimization in Post-Training Quantization for Large Language Models View paper
- [50] Scaling Laws for Task-Stratified Knowledge in Post-Training Quantized Large Language Models View paper
- [51] OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting View paper

- [52] Neural Networks with Model Compression View paper
- [53] Color conversion matrices in digital cameras: a tutorial View paper
- [54] Mixture attention block and Swin transformerâ□□based entropy model for learned image compression View paper
- [55] A Comprehensive Evaluation on Quantization Techniques for Large Language Models View paper
- [56] Systematic codebook designs for quantized beamforming in correlated MIMO channels View paper
- [57] Deep learning image compression with multi-channel tANS coding and hardware deployment View paper
- [58] Quantization Methods for Matrix Multiplication and Efficient Transformers View paper
- [59] BASE-Q: Bias and Asymmetric Scaling Enhanced Rotational Quantization for Large Language Models View paper
- [60] BTC-LLM: Efficient Sub-1-Bit LLM Quantization via Learnable Transformation and Binary Codebook View paper
- [61] Quarot: Outlier-free 4-bit inference in rotated llms View paper
- [62] Duquant: Distributing outliers via dual transformation makes stronger quantized llms View paper
- [63] Turning LLM Activations Quantization-Friendly View paper
- [64] SmoothRot: Combining Channel-Wise Scaling and Rotation for Quantization-Friendly LLMs View paper
- [65] Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations View paper
- [66] Quantized Visual Geometry Grounded Transformer View paper
- [67] Rotated Runtime Smooth: Training-Free Activation Smoother for accurate INT4 inference View paper
- [68] Rolora: Fine-tuning rotated outlier-free llms for effective weight-activation quantization View paper
- [69] Rotate, Clip, and Partition: Towards W2A4KV4 Quantization by Integrating Rotation and Learnable Non-uniform Quantizer View paper
- [70] TR-DQ: Time-Rotation Diffusion Quantization View paper
- [71] ConvRot: Rotation-Based Plug-and-Play 4-bit Quantization for Diffusion Transformers View paper
- [72] Bridging the Gap Between Promise and Performance for Microscaling FP4 Quantization View paper
- [73] KVLinC: KV Cache Quantization with Hadamard Rotation and Linear Correction View paper
- [74] ROSAQ: Rotation-based Saliency-Aware Weight Quantization for Efficiently Compressing Large Language Models View paper
- [75] Q-Palette: Fractional-Bit Quantizers Toward Optimal Bit Allocation for Efficient LLM Deployment View paper
- [76] Pushing the Limits of Large Language Model Quantization via the Linearity Theorem View paper
- [77] Breaking the Efficiency-Accuracy: Fusion of Rotation Quantization and N: M Sparsity for LLMs Inference View paper