Novelty Assessment Report

Paper: QWHA: Quantization-Aware Walsh-Hadamard Adaptation for Parameter-Efficient Fine-Tuning on Large Language Models

PDF URL: https://openreview.net/pdf?id=QMN4ERDdp4

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

The demand for efficient deployment of large language models (LLMs) has driven interest in quantization, which reduces inference cost, and parameter-efficient fine-tuning (PEFT), which lowers training overhead. This motivated the development of quantization-aware PEFT to produce accurate yet efficient quantized models. In this setting, reducing quantization error prior to fine-tuning is crucial for achieving high model accuracy. However, existing methods that rely on low-rank adaptation suffer from limited representational capacity. Recent Fourier-related transform (FT)-based adapters offer greater representational power than low-rank adapters, but their direct integration into quantized models often results in ineffective error reduction and increased computational overhead.

To overcome these limitations, we propose QWHA, a method that integrates FT-based adapters into quantized models by employing the Walsh-Hadamard Transform (WHT) as the transform kernel, together with a novel adapter initialization scheme incorporating adaptive parameter selection and value refinement. We demonstrate that QWHA effectively mitigates quantization errors while facilitating fine-tuning, and that its design substantially reduces computational cost. Experimental results show that QWHA consistently outperforms baselines in low-bit quantization accuracy and achieves significant training speedups over existing FT-based adapters.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Quantization-Aware Parameter-Efficient Fine-Tuning for Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Quantization-Aware Training Frameworks**
- **Outlier and Activation Management**
- **Specialized Quantization Techniques**
- **Task-Agnostic and Pre-Training Quantization**
- **Efficient Deployment and Inference Optimization**
- **Domain-Specific and Multimodal Applications**
- **Reinforcement Learning and Advanced Training Paradigms**
- **Comprehensive Evaluation and Benchmarking**
- **Survey and Review Literature**
- **Foundational PEFT and Quantization Methods**

Complete Taxonomy Tree

- Quantization-Aware Parameter-Efficient Fine-Tuning for Large Language Models Survey Taxonomy
- Quantization-Aware Training Frameworks
  - Full Quantization-Aware Training (4 papers)
  - [1] Llm-qat: Data-free quantization aware training for large language models (Zechun Liu, 2024) View paper
  - [4] Efficientqat: Efficient quantization-aware training for large language models (Mengzhao Chen, 2025) View paper
  - [16] RoSTE: An Efficient Quantization-Aware Supervised Fine-Tuning Approach for Large Language Models (Wei Quan, 2025) View paper
  - [37] SASQ: Static Activation Scaling for Quantization-Aware Training in Large Language Models (Shizhuo Mao, 2025) View paper
  - Low-Rank Adaptation with Quantization Awareness
  - Group-Wise and Decomposed QAT-LoRA (3 papers)
    - [2] Qa-lora: Quantization-aware low-rank adaptation of large language models (XU Yuhui, 2023) View paper
    - [7] Dl-qat: Weight-decomposed low-rank quantization-aware training for large language models (KE Wenjin, 2024) View paper
    - [17] Low-Rank Quantization-Aware Training for LLMs (Bondarenko, 2024) View paper
  - Initialization-Aware Quantization for LoRA (3 papers)
    - [5] Loftq: Lora-fine-tuning-aware quantization for large language models (Li, 2023) View paper
    - [8] Qeft: Quantization for efficient fine-tuning of llms (Chang-Hun Lee, 2024) View paper
    - [9] Accurate and efficient fine-tuning of quantized large language models through optimal balance (Shen Ao, 2024) View paper
  - Rank-Adaptive and Dynamic Precision LoRA (3 papers)
    - [13] RA-LoRA: Rank-Adaptive Parameter-Efficient Fine-Tuning for Accurate 2-bit Quantized Large Language Models (Minsoo Kim, 2024) View paper
    - [31] Efficient fine-tuning of quantized models via adaptive rank and bitwidth (Zhou Chang-hai, 2025) View paper
    - [38] On-the-Fly Adaptation to Quantization: Configuration-Aware LoRA for Efficient Fine-Tuning of Quantized LLMs (Tang Ming, 2025) View paper
  - Transform-Based Quantization-Aware Adaptation ★ (2 papers)

  - [0] QWHA: Quantization-Aware Walsh-Hadamard Adaptation for Parameter-Efficient Fine-Tuning on Large Language Models (Anon et al., 2026) View paper
  - [50] HALO: Hadamard-Assisted Lower-Precision Optimization for LLMs (Ashkboos, 2025) View paper
- Outlier and Activation Management
  - Outlier-Aware Mixed-Precision Quantization (2 papers)
  - [3] Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models (Changhun Lee, 2024) View paper
  - [47] MoQAE: Mixed-Precision Quantization for Long-Context LLM Inference via Mixture of Quantization-Aware Experts (Wei Tao, 2025) View paper
  - Activation Outlier Suppression via Training (4 papers)
  - [12] Taming sensitive weights: Noise perturbation fine-tuning for robust llm quantization (Wang Dongwei, 2024) View paper
  - [25] RoLoRA: Fine-tuning Rotated Outlier-free LLMs for Effective Weight-Activation Quantization (Xijie Huang, 2024) View paper
  - [29] QUAD: Quantization and Parameter-Efficient Tuning of LLM with Activation Decomposition (Hu, 2025) View paper
  - [36] BiSup: Bidirectional Quantization Error Suppression for Large Language Models (Zou Ming-hui, 2024) View paper
- Specialized Quantization Techniques
  - Sub-4-Bit and Extreme Low-Bit Quantization (2 papers)
  - [14] Memory-Efficient Fine-Tuning of Compressed Large Language Models via sub-4-bit Integer Quantization (Kim Jeonghoon, 2023) View paper
  - [46] PB-LLM: Partially Binarized Large Language Models (Shang, 2023) View paper
  - Floating-Point and Alternative Quantization Formats (1 papers)
  - [33] FP4-Quantization: Lossless 4bit Quantization for Large Language Models (Jie Wang, 2024) View paper
- Task-Agnostic and Pre-Training Quantization (1 papers)
  - [18] PreQuant: A Task-agnostic Quantization Approach for Pre-trained Language Models (Zhuocheng Gong, 2023) View paper
- Efficient Deployment and Inference Optimization
  - Multi-Configuration and Adaptive Deployment (3 papers)
  - [10] L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models (HyeSung Jeon, 2024) View paper
  - [11] L4Q: Parameter Efficient Quantization-Aware Training on Large Language Models via LoRA-wise LSQ (Jeon, 2024) View paper
  - [22] One quantllm for all: Fine-tuning quantized llms once for efficient deployments (Ke Yi, 2025) View paper
  - Side Tuning and Memory-Efficient Inference (1 papers)
  - [28] Quantized side tuning: Fast and memory-efficient tuning of quantized large language models (Zhengxin Zhang, 2024) View paper
  - Hardware-Aware Quantization Processors (1 papers)
  - [42] A 28nm 3.14 TFLOP/W BF16 LLM Fine-Tuning Processor with Asymmetric Quantization Computing for AI PC (Xinyuan Lin, 2025) View paper
- Domain-Specific and Multimodal Applications
  - Multimodal Large Language Models (2 papers)
  - [15] Advancing Multimodal Large Language Models with Quantization-Aware Scale Learning for Efficient Adaptation (Xie Jing-Jing, 2024) View paper
  - [27] AndesVL Technical Report: An Efficient Mobile-side Multimodal Large Language Model (Jin Zhiwei, 2025) View paper
  - Domain-Specific Fine-Tuning Applications (7 papers)
  - [21] Fine-Tuning an LLM Using QLORA and PEFT with RLHF Dataset (T Grace Shalini, 2025) View paper
  - [24] Multilevel Analysis of Cryptocurrency News using RAG Approach with Fine-Tuned Mistral Large Language Model (Pavlyshenko, 2025) View paper
  - [30] Enhancing Medical Summarization with Parameter Efficient Fine Tuning on Local CPUs (Goh Man Fye, 2024) View paper
  - [32] NEFMind: Parameter-Efficient Fine-Tuning of Open-Source LLMs for Telecom APIs Automation (Hussain Ahmed, 2025) View paper
  - [39] Finetuning Open-Source LLMs for Teaching Purpose: Efficient Parameter-Efficient Fine-Tuning Using QLoRA on Consumer Hardware (Saxena, 2025) View paper
  - [41] Domain Specific Finetuning of LLMs Using PEFT Techniques (Deva Kumar Gajulamandyam, 2025) View paper
  - [45] Industrial Feasibility of PEFTpyCoder: A Parameter-Efficient Fine-Tuned Model for In-House Code Assistance (Adnan Riaz, 2025) View paper
- Reinforcement Learning and Advanced Training Paradigms (1 papers)
  - [44] QeRL: Beyond Efficiency--Quantization-enhanced Reinforcement Learning for LLMs (Huang Wei, 2025) View paper
- Comprehensive Evaluation and Benchmarking (3 papers)
  - [26] Optimizing Fine-Tuning in Quantized Language Models: An In-Depth Analysis of Key Variables. (Ao Shen, 2025) View paper
  - [43] Accurate and Efficient Fine-Tuning of Quantized Large Language Models Through Optimal Balance in Adaptation (Ao Shen, 2025) View paper
  - [49] EfficientLLM: Efficiency in Large Language Models (Yuan, 2025) View paper
- Survey and Review Literature (5 papers)
  - [19] A survey on 1-bit quantized large language models (Kritika Tripathi, 2025) View paper
  - [20] Exploring quantization techniques for large-scale language models: Methods, challenges and future directions (Ao Shen, 2024) View paper
  - [23] Efficient compressing and tuning methods for large language models: A systematic literature review (Gun Il Kim, 2025) View paper
  - [34] A survey of low-bit large language models: Basics, systems, and algorithms (Ruihao Gong, 2024) View paper
  - [35] llm fine-tuning: Instruction and parameterefficient fine-tuning (peft) (S Aathilakshmi, 2024) View paper
- Foundational PEFT and Quantization Methods (3 papers)
  - [6] Alphatuning: Quantization-aware parameter-efficient adaptation of large-scale pre-trained language models (Se Jung Kwon, 2022) View paper
  - [40] A Quantization Approach for the Reduced Size of Large Language Models (Ravi Kishore Kodali, 2024) View paper
  - [48] QLoRA: Efficient Finetuning of Quantized LLMs (Dettmers, 2023) View paper

Narrative

Core task: quantization-aware parameter-efficient fine-tuning for large language models. This field addresses the challenge of adapting massive pre-trained models to downstream tasks while simultaneously reducing memory and computational costs through quantization. The taxonomy reveals a rich landscape organized around several complementary themes. Quantization-Aware Training Frameworks explore methods that integrate low-bit representations directly into the fine-tuning loop, often combining techniques like low-rank adaptation with learned quantization parameters (e.g., QA-LoRA[2], LLM-QAT[1]). Outlier and Activation Management tackles the problem of extreme values that degrade quantized model quality, while Specialized Quantization Techniques investigate novel bit-width schemes and mixed-precision strategies. Task-Agnostic and Pre-Training Quantization focuses on compressing models before task-specific adaptation, and Efficient Deployment branches emphasize inference-time optimizations. Domain-Specific and Multimodal Applications extend these ideas to specialized settings, and Reinforcement Learning paradigms incorporate quantization into policy optimization. Comprehensive Evaluation and Benchmarking provides systematic comparisons, while Survey and Review Literature synthesizes progress across the field. Foundational PEFT and Quantization Methods anchor the taxonomy with seminal works like QLoRA[48] and LoftQ[5].

A particularly active line of work centers on transform-based and initialization-aware strategies that carefully balance quantization error with adaptation capacity. QWHA[0] exemplifies this direction by proposing a Hadamard-transform approach to mitigate quantization-induced distortions during parameter-efficient tuning, situating itself within the Quantization-Aware Training Frameworks branch alongside methods like HALO[50] that similarly manipulate weight representations before quantization. This contrasts with approaches such as LoftQ[5], which emphasizes joint initialization of quantized weights and low-rank adapters, or OWQ[3], which prioritizes outlier-aware schemes. The interplay between these strategies highlights a central trade-off: whether to invest effort in pre-quantization transformations, adaptive rank selection, or outlier suppression. QWHA[0] leans toward transformation-based mitigation, offering a complementary perspective to works that adjust rank dynamically or manage activations explicitly, and reflects ongoing exploration of how structural interventions can preserve fine-tuning expressiveness under aggressive compression.

Related Works in Same Category

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

1. HALO: Hadamard-Assisted Lower-Precision Optimization for LLMs

Authors: Ashkboos, Saleh, Nikdan, Mahdi, Saleh Ashkboos, et al. (18 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Quantized training of Large Language Models (LLMs) remains an open challenge, as maintaining accuracy while performing all matrix multiplications in low precision has proven difficult. This is particularly the case when fine-tuning pre-trained models, which can have large weight and activation outlier values that make lower-precision optimization difficult. To address this, we present HALO, a novel quantization-aware training approach for Transformers that enables accurate and efficient low-prec...

Relationship Analysis

Both papers belong to the Transform-Based Quantization-Aware Adaptation category, employing orthogonal transforms (Walsh-Hadamard Transform) to enhance representational capacity in quantized parameter-efficient fine-tuning. They overlap in using WHT to mitigate quantization errors and improve low-precision training of LLMs. However, QWHA focuses on adapter initialization strategies (AdaAlloc and Refinement) for quantization-aware PEFT with sparse coefficient matrices, while HALO emphasizes strategic placement of Hadamard rotations in both forward and backward passes for full fine-tuning with FSDP integration and quantized communication.

Contributions Analysis

Overall novelty summary. The paper proposes QWHA, a method that integrates Walsh-Hadamard Transform-based adapters into quantized large language models with a novel initialization scheme (AdaAlloc) for adaptive parameter selection and value refinement. Within the taxonomy, it resides in the 'Transform-Based Quantization-Aware Adaptation' leaf under 'Quantization-Aware Training Frameworks'. This leaf contains only two papers total, indicating a relatively sparse research direction compared to more crowded areas like 'Low-Rank Adaptation with Quantization Awareness', which spans multiple sub-categories with numerous papers.

The taxonomy structure reveals that QWHA's immediate neighbors include low-rank adaptation methods (LoRA variants with group-wise quantization, initialization-aware schemes, and rank-adaptive approaches) and outlier management techniques. The 'Transform-Based' leaf sits alongside these more populated branches, suggesting that Fourier-related and orthogonal transform approaches represent an emerging alternative to dominant low-rank paradigms. The taxonomy's scope note explicitly distinguishes transform-based methods from rotation-based outlier suppression and pure low-rank techniques, positioning QWHA at the intersection of representational capacity enhancement and quantization error mitigation.

Among 30 candidates examined, the contribution-level analysis shows mixed novelty signals. The overall QWHA method (10 candidates examined, 0 refutable) and the AdaAlloc initialization scheme (10 candidates examined, 0 refutable) appear to have no clear prior work overlap within the limited search scope. However, the Walsh-Hadamard Transform-based adapter design itself (10 candidates examined, 1 refutable) shows at least one candidate providing overlapping prior work. This suggests the transform kernel choice may have precedent, while the integration strategy and initialization approach appear more distinctive within the examined literature.

Based on the limited top-30 semantic search scope, QWHA appears to occupy a relatively under-explored niche combining transform-based adaptation with quantization-aware initialization. The sparse population of its taxonomy leaf and the lack of refutable candidates for two of three contributions suggest potential novelty, though the single refutable candidate for the WHT adapter design indicates some prior exploration of transform kernels. A more exhaustive search would be needed to definitively assess originality across the broader quantization-aware PEFT landscape.

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

Contribution 1: QWHA method integrating Walsh-Hadamard Transform-based adapter with quantization-aware initialization

Description: The authors introduce QWHA, which combines a Walsh-Hadamard Transform-based adapter (WHA) with a quantization-aware initialization strategy for parameter-efficient fine-tuning of quantized large language models. This method addresses limitations of existing low-rank and Fourier-transform-based adapters in the quantization-aware setting.

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. LoRA: Low-Rank Adaptation of Large Language Models

URL: View paper

Brief Assessment

LoRA[54] introduces low-rank adaptation for parameter-efficient fine-tuning but does not address Walsh-Hadamard Transform-based adapters or quantization-aware initialization strategies. The candidate focuses on rank decomposition matrices without transform-based methods or quantization considerations.

2. Dynamic Low-Rank Sparse Adaptation for Large Language Models

URL: View paper

Brief Assessment

Dynamic Low-Rank Sparse[57] focuses on sparsity-based adaptation for LLMs, not quantization-aware fine-tuning with transform-based adapters. The candidate addresses sparse model fine-tuning while the original addresses quantized model adaptation.

3. LoRS: Efficient Low-Rank Adaptation for Sparse Large Language Model

URL: View paper

Brief Assessment

LoRS[53] focuses on maintaining sparsity in sparse LLMs through memory-efficient LoRA adaptations, not on Walsh-Hadamard Transform-based adapters or quantization-aware initialization for quantized models.

4. Lost: Low-rank and sparse pre-training for large language models

URL: View paper

Brief Assessment

Lost[55] focuses on low-rank and sparse pre-training for LLMs from scratch, not on parameter-efficient fine-tuning of quantized models with transform-based adapters. The technical approaches and problem settings are fundamentally different.

5. DenseLoRA: Dense Low-Rank Adaptation of Large Language Models

URL: View paper

Brief Assessment

DenseLoRA[56] focuses on improving parameter efficiency in standard (non-quantized) fine-tuning through representation refinement and dense low-rank matrices. It does not address quantization-aware initialization or transform-based adapters for quantized models, which are the core technical contributions of the original paper.

6. Cora: Optimizing low-rank adaptation with common subspace of large language models

URL: View paper

Brief Assessment

Cora[52] focuses on optimizing LoRA by extracting a common subspace matrix from multiple fine-tuned models to replace the B matrix in LoRA, not on quantization-aware parameter-efficient fine-tuning or transform-based adapters for quantized models.

7. Loftq: Lora-fine-tuning-aware quantization for large language models

URL: View paper

Brief Assessment

LoftQ[5] focuses on quantization-aware initialization for LoRA adapters using SVD-based low-rank decomposition, not transform-based adapters like Walsh-Hadamard Transform. The methods address different technical approaches to quantization-aware parameter-efficient fine-tuning.

8. QLoRA: Efficient Finetuning of Quantized LLMs

URL: View paper

Brief Assessment

QLoRA[48] focuses on efficient finetuning of quantized LLMs using LoRA adapters with 4-bit quantization techniques (NF4, double quantization), not on Walsh-Hadamard Transform-based adapters or quantization-aware initialization strategies for parameter selection.

9. Sparse low-rank adaptation of pre-trained language models

URL: View paper

Brief Assessment

Sparse Low-Rank[51] focuses on dynamic rank adaptation in LoRA through sparsity-inducing gates for standard PEFT, not on quantization-aware initialization or transform-based adapters for quantized models.

10. Qa-lora: Quantization-aware low-rank adaptation of large language models

URL: View paper

Brief Assessment

QA-LoRA[2] focuses on group-wise quantization with low-rank adaptation (LoRA) for efficient fine-tuning, not on Walsh-Hadamard Transform-based adapters or Fourier-transform methods. The technical approaches are fundamentally different.

Contribution 2: Walsh-Hadamard Transform-based adapter (WHA) design

Description: The authors design a novel adapter using the Walsh-Hadamard Transform as the transform kernel, which consists only of ±1 entries enabling efficient computation through additions and subtractions. Unlike conventional Fourier-transform-based adapters, WHA applies a single transform rather than double transforms, reducing computational overhead while maintaining superior representational capacity.

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. A 22nm 9.51 TOPS/W Neural Engine with 2MB MRAM Leveraging Sparse-Orthogonal Walsh-Hadamard Transform Computations and Dynamic Power Gating

URL: View paper

Brief Assessment

Walsh-Hadamard Neural Engine[68] focuses on hardware implementation of Walsh-Hadamard transform for neural network processors with MRAM storage, not on adapter design for parameter-efficient fine-tuning of language models.

2. Block Walshâ□□Hadamard Transform-based Binary Layers in Deep Neural Networks

URL: View paper

Brief Assessment

Block Walsh-Hadamard Binary[71] focuses on replacing convolutional layers in CNNs for image classification, not on adapter design for parameter-efficient fine-tuning of LLMs. The technical contexts are fundamentally different: one addresses network compression in computer vision, the other addresses efficient adaptation in language models.

3. SuperLoRA: Parameter-Efficient Unified Adaptation of Large Foundation Models

URL: View paper

Brief Assessment

SuperLoRA[76] focuses on unifying various LoRA variants through grouping, folding, and projection mechanisms for parameter-efficient fine-tuning. While it mentions Walsh-Hadamard transforms in the context of fastfood projection (a fixed projection layer), it does not propose a Walsh-Hadamard Transform-based adapter as the core adaptation mechanism. The original paper's WHA uses WHT as the primary transform kernel for weight updates, whereas SuperLoRA[76] uses WHT only within an optional projection function alongside other techniques.

4. Walsh-hadamard variational inference for bayesian deep learning

URL: View paper

Brief Assessment

Walsh-Hadamard Variational Inference[75] applies Walsh-Hadamard transforms to variational inference for Bayesian deep learning, not to adapter design for parameter-efficient fine-tuning. The candidate focuses on approximating posterior distributions in Bayesian neural networks, while the original paper designs adapters for efficient weight updates in quantized LLM fine-tuning.

5. Unleashing the Power of Random Projection for Efficient Machine Learning

URL: View paper

Brief Assessment

Random Projection[77] focuses on dimensionality reduction and kernel approximation using Walsh-Hadamard Transform for data compression and efficient machine learning, not on adapter design for neural network fine-tuning or parameter-efficient updates in large language models.

6. Voltage Based Electronic Control Unit (ECU) Identification with Convolutional Neural Networks and Walshâ□□Hadamard Transform

URL: View paper

Brief Assessment

ECU Identification Walsh-Hadamard[72] applies Walsh-Hadamard Transform to ECU voltage fingerprint identification using CNNs, not to parameter-efficient fine-tuning of language models. The application domains and technical contexts are entirely different.

7. Fast Walsh-Hadamard Transform and Smooth-Thresholding Based Binary Layers in Deep Neural Networks

URL: View paper

Brief Assessment

Fast Walsh-Hadamard Binary[74] focuses on replacing 1x1 convolution layers in CNNs with FWHT layers for image classification tasks, not on parameter-efficient fine-tuning of LLMs with adapters for weight updates as in the original paper.

8. Fast randomized low-rank adaptation of pre-trained language models with pac regularization

URL: View paper

Prior Art Analysis

Fast Randomized LoRA[69] demonstrates prior work that uses Walsh-Hadamard Transform (WHT) for efficient adapter design in parameter-efficient fine-tuning. The candidate paper explicitly describes using randomized WHT to construct low-rank projection matrices, replacing the trainable matrices in LoRA with WHT-based approximations. This approach predates the original paper's WHA design and shares the core innovation of leveraging WHT's ±1 structure for computational efficiency through additions and subtractions rather than matrix multiplications. Both papers exploit the same fundamental property: WHT's recursive definition enabling O(d log d) computation without explicit matrix storage.

Evidence

Evidence 1 - **Rationale**: Fast Randomized LoRA[69] explicitly describes the same recursive WHT structure with ±1 entries and O(d log d) computation that the original paper claims as novel. - **Original**: the wht kernel consists solely of ±1 elements, enabling efficient computations using only additions and subtractions, thereby eliminating matrix multiplications - **Candidate**: the matrix h ∈rdxd, it can be efficiently generated due to its recursive definition: hd =[hd/2 hd/2 hd/2 -hd/2] with h2 = [1 1 1 -1] . the generation iterates two simple steps: i) dividing the data into two halves, and ii) performing a simple ±operation. this fast wht allows hx to be computed in o...

9. Real-Time Low-Cost Drift Compensation for Chemical Sensors Using a Deep Neural Network With Hadamard Transform and Additive Layers

URL: View paper

Brief Assessment

Chemical Sensors Hadamard[70] applies Hadamard transforms to chemical sensor drift compensation in a completely different domain (sensor signal processing), not to neural network weight adaptation or parameter-efficient fine-tuning of language models.

10. Efficient Transformations in Deep Learning Convolutional Neural Networks

URL: View paper

Brief Assessment

Efficient Transformations[73] applies Walsh-Hadamard Transform to convolutional neural networks for image classification, not to parameter-efficient fine-tuning adapters for large language models. The technical contexts are fundamentally different.

Contribution 3: AdaAlloc parameter selection and value refinement initialization scheme

Description: The authors develop a tractable initialization solution consisting of AdaAlloc, which adaptively allocates parameters across output channels proportional to their quantization errors while ensuring full-rank capacity, and a value refinement step that re-projects selected parameters to minimize layer output error. This initialization effectively reduces quantization errors before fine-tuning.

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. Plug-and-play 1. x-bit kv cache quantization for video large language models

URL: View paper

Brief Assessment

KV Cache Quantization[58] focuses on quantizing key-value caches in video large language models for inference efficiency, not on adaptive parameter allocation for channel-wise neural network compression during training or fine-tuning.

2. Adaptive quantization and pruning of deep neural networks via layer importance estimation

URL: View paper

Brief Assessment

Layer Importance Estimation[61] focuses on adaptive quantization and pruning for model compression, not on parameter-efficient fine-tuning initialization schemes for quantized models. The candidate addresses layer-wise bit-width allocation and pruning thresholds, while the original contribution concerns adapter parameter selection and value refinement for reducing quantization errors before fine-tuning.

3. Joint optimization of dimension reduction and mixed-precision quantization for activation compression of neural networks

URL: View paper

Brief Assessment

Dimension Reduction Mixed-Precision[62] focuses on joint optimization of dimension reduction and mixed-precision quantization for activation compression, not on parameter-efficient fine-tuning with adaptive parameter allocation for quantization error reduction in the context of adapters for large language models.

4. Distribution Adaptive INT8 Quantization for Training CNNs

URL: View paper

Brief Assessment

Distribution Adaptive INT8[66] focuses on gradient quantization for CNN training acceleration, not on parameter-efficient fine-tuning adapter initialization. The candidate's channel-wise gradient quantization addresses training stability in low-precision arithmetic, while the original paper's AdaAlloc allocates adapter parameters to minimize quantization errors before fine-tuning LLMs—fundamentally different problem domains and technical approaches.

5. Towards optimal layer ordering for efficient model compression via pruning and quantization

URL: View paper

Brief Assessment

Optimal Layer Ordering[63] focuses on layer sequencing for compression optimization in DNNs, not on adaptive parameter allocation schemes for reducing quantization errors in channel-wise neural network compression as proposed in the original paper's AdaAlloc method.

6. A Survey on Network Quantization Techniques for Deep Neural Network Compression

URL: View paper

Brief Assessment

Network Quantization Survey[65] is a survey paper discussing general quantization techniques. The candidate's brief mentions of 'adaptive approaches' and 'layer-wise loss balancing' do not specifically address the AdaAlloc scheme's novel combination of channel-wise parameter allocation proportional to quantization errors with full-rank capacity guarantees, nor the value refinement re-projection step for minimizing layer output error in the context of quantization-aware fine-tuning with transform-based adapters.

7. On-Device Large Language Models: A Survey of Model Compression and System Optimization

URL: View paper

Brief Assessment

On-Device LLM Survey[64] is a broad survey paper covering model compression techniques including quantization, pruning, and low-rank methods. It does not present a specific adaptive parameter allocation scheme for channel-wise quantization error reduction comparable to AdaAlloc.

8. Distribution-aware adaptive multi-bit quantization

URL: View paper

Brief Assessment

Distribution-aware Multi-bit[67] focuses on multi-bit quantization for neural network compression with distribution-aware quantization schemes and loss-guided bit-width allocation. The candidate does not address parameter-efficient fine-tuning adapters or the specific initialization challenges in quantization-aware PEFT contexts that the original paper tackles.

9. Channel-wise mixed-precision quantization for large language models

URL: View paper

Brief Assessment

Channel-wise Mixed-Precision[59] focuses on mixed-precision quantization for LLMs by allocating different bit-widths to weight channels based on activation distributions, not on adaptive parameter allocation for reducing quantization errors in the context of parameter-efficient fine-tuning adapters.

10. Bayesian bits: Unifying quantization and pruning

URL: View paper

Brief Assessment

Bayesian Bits[60] focuses on learning mixed-precision quantization through stochastic gates and residual decomposition for power-of-two bit widths. It does not address adaptive parameter allocation across output channels based on quantization errors or value refinement for channel-wise error minimization, which are the core innovations of AdaAlloc.

Appendix: Text Similarity Detection

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

References

- [0] QWHA: Quantization-Aware Walsh-Hadamard Adaptation for Parameter-Efficient Fine-Tuning on Large Language Models View paper
- [1] Llm-qat: Data-free quantization aware training for large language models View paper
- [2] Qa-lora: Quantization-aware low-rank adaptation of large language models View paper

- [3] Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models View paper
- [4] Efficientqat: Efficient quantization-aware training for large language models View paper
- [5] Loftq: Lora-fine-tuning-aware quantization for large language models View paper
- [6] Alphatuning: Quantization-aware parameter-efficient adaptation of large-scale pre-trained language models View paper
- [7] Dl-qat: Weight-decomposed low-rank quantization-aware training for large language models View paper
- [8] Qeft: Quantization for efficient fine-tuning of llms View paper
- [9] Accurate and efficient fine-tuning of quantized large language models through optimal balance View paper
- [10] L4Q: Parameter Efficient Quantization-Aware Fine-Tuning on Large Language Models View paper
- [11] L4Q: Parameter Efficient Quantization-Aware Training on Large Language Models via LoRA-wise LSQ View paper
- [12] Taming sensitive weights: Noise perturbation fine-tuning for robust llm quantization View paper
- [13] RA-LoRA: Rank-Adaptive Parameter-Efficient Fine-Tuning for Accurate 2-bit Quantized Large Language Models View paper
- [14] Memory-Efficient Fine-Tuning of Compressed Large Language Models via sub-4-bit Integer Quantization View paper
- [15] Advancing Multimodal Large Language Models with Quantization-Aware Scale Learning for Efficient Adaptation View paper
- [16] RoSTE: An Efficient Quantization-Aware Supervised Fine-Tuning Approach for Large Language Models View paper
- [17] Low-Rank Quantization-Aware Training for LLMs View paper
- [18] PreQuant: A Task-agnostic Quantization Approach for Pre-trained Language Models View paper
- [19] A survey on 1-bit quantized large language models View paper
- [20] Exploring quantization techniques for large-scale language models: Methods, challenges and future directions View paper
- [21] Fine-Tuning an LLM Using QLORA and PEFT with RLHF Dataset View paper
- [22] One quantllm for all: Fine-tuning quantized llms once for efficient deployments View paper
- [23] Efficient compressing and tuning methods for large language models: A systematic literature review View paper
- [24] Multilevel Analysis of Cryptocurrency News using RAG Approach with Fine-Tuned Mistral Large Language Model View paper
- [25] RoLoRA: Fine-tuning Rotated Outlier-free LLMs for Effective Weight-Activation Quantization View paper
- [26] Optimizing Fine-Tuning in Quantized Language Models: An In-Depth Analysis of Key Variables. View paper
- [27] AndesVL Technical Report: An Efficient Mobile-side Multimodal Large Language Model View paper
- [28] Quantized side tuning: Fast and memory-efficient tuning of quantized large language models View paper
- [29] QUAD: Quantization and Parameter-Efficient Tuning of LLM with Activation Decomposition View paper
- [30] Enhancing Medical Summarization with Parameter Efficient Fine Tuning on Local CPUs View paper
- [31] Efficient fine-tuning of quantized models via adaptive rank and bitwidth View paper
- [32] NEFMind: Parameter-Efficient Fine-Tuning of Open-Source LLMs for Telecom APIs Automation View paper
- [33] FP4-Quantization: Lossless 4bit Quantization for Large Language Models View paper
- [34] A survey of low-bit large language models: Basics, systems, and algorithms View paper
- [35] llm fine-tuning: Instruction and parameterefficient fine-tuning (peft) View paper
- [36] BiSup: Bidirectional Quantization Error Suppression for Large Language Models View paper
- [37] SASQ: Static Activation Scaling for Quantization-Aware Training in Large Language Models View paper
- [38] On-the-Fly Adaptation to Quantization: Configuration-Aware LoRA for Efficient Fine-Tuning of Quantized LLMs View paper
- [39] Finetuning Open-Source LLMs for Teaching Purpose: Efficient Parameter-Efficient Fine-Tuning Using QLoRA on Consumer Hardware View paper
- [40] A Quantization Approach for the Reduced Size of Large Language Models View paper
- [41] Domain Specific Finetuning of LLMs Using PEFT Techniques View paper
- [42] A 28nm 3.14 TFLOP/W BF16 LLM Fine-Tuning Processor with Asymmetric Quantization Computing for AI PC View paper
- [43] Accurate and Efficient Fine-Tuning of Quantized Large Language Models Through Optimal Balance in Adaptation View paper
- [44] QeRL: Beyond Efficiency--Quantization-enhanced Reinforcement Learning for LLMs View paper
- [45] Industrial Feasibility of PEFTpyCoder: A Parameter-Efficient Fine-Tuned Model for In-House Code Assistance View paper
- [46] PB-LLM: Partially Binarized Large Language Models View paper
- [47] MoQAE: Mixed-Precision Quantization for Long-Context LLM Inference via Mixture of Quantization-Aware Experts View paper
- [48] QLoRA: Efficient Finetuning of Quantized LLMs View paper
- [49] EfficientLLM: Efficiency in Large Language Models View paper
- [50] HALO: Hadamard-Assisted Lower-Precision Optimization for LLMs View paper
- [51] Sparse low-rank adaptation of pre-trained language models View paper
- [52] Cora: Optimizing low-rank adaptation with common subspace of large language models View paper
- [53] LoRS: Efficient Low-Rank Adaptation for Sparse Large Language Model View paper
- [54] LoRA: Low-Rank Adaptation of Large Language Models View paper
- [55] Lost: Low-rank and sparse pre-training for large language models View paper
- [56] DenseLoRA: Dense Low-Rank Adaptation of Large Language Models View paper
- [57] Dynamic Low-Rank Sparse Adaptation for Large Language Models View paper
- [58] Plug-and-play 1. x-bit kv cache quantization for video large language models View paper
- [59] Channel-wise mixed-precision quantization for large language models View paper
- [60] Bayesian bits: Unifying quantization and pruning View paper
- [61] Adaptive quantization and pruning of deep neural networks via layer importance estimation View paper
- [62] Joint optimization of dimension reduction and mixed-precision quantization for activation compression of neural networks View paper
- [63] Towards optimal layer ordering for efficient model compression via pruning and quantization View paper
- [64] On-Device Large Language Models: A Survey of Model Compression and System Optimization View paper
- [65] A Survey on Network Quantization Techniques for Deep Neural Network Compression View paper
- [66] Distribution Adaptive INT8 Quantization for Training CNNs View paper
- [67] Distribution-aware adaptive multi-bit quantization View paper
- [68] A 22nm 9.51 TOPS/W Neural Engine with 2MB MRAM Leveraging Sparse-Orthogonal Walsh-Hadamard Transform Computations and Dynamic Power Gating View paper
- [69] Fast randomized low-rank adaptation of pre-trained language models with pac regularization View paper
- [70] Real-Time Low-Cost Drift Compensation for Chemical Sensors Using a Deep Neural Network With Hadamard Transform and Additive Layers View paper

- [71] Block Walshâ□□Hadamard Transform-based Binary Layers in Deep Neural Networks View paper
- [72] Voltage Based Electronic Control Unit (ECU) Identification with Convolutional Neural Networks and Walshâ□□Hadamard Transform View paper
- [73] Efficient Transformations in Deep Learning Convolutional Neural Networks View paper
- [74] Fast Walsh-Hadamard Transform and Smooth-Thresholding Based Binary Layers in Deep Neural Networks View paper
- [75] Walsh-hadamard variational inference for bayesian deep learning View paper
- [76] SuperLoRA: Parameter-Efficient Unified Adaptation of Large Foundation Models View paper
- [77] Unleashing the Power of Random Projection for Efficient Machine Learning View paper