Novelty Assessment Report

Paper: ARMOR: High-Performance Semi-Structured Pruning via Adaptive Matrix Factorization
PDF URL: https://openreview.net/pdf?id=8NE554wv0m
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Large language models (LLMs) present significant deployment challenges due to their immense computational and memory requirements. While semi-structured pruning, particularly 2:4 sparsity, offers a path to practical hardware acceleration, existing methods often incur substantial performance degradation. To bridge this gap, we introduce ARMOR: (Adaptive Representation with Matrix-factORization), a novel one-shot post-training pruning algorithm. Instead of directly pruning weights, ARMOR factorizes each weight matrix into a 2:4 sparse core wrapped by two low-overhead, block diagonal matrices. These wrappers act as efficient pre- and post-transformation error correctors, offering greater flexibility to preserve model quality compared to conventional 2:4 pruning techniques. The sparse core and block diagonal wrappers are chosen through a block coordinate descent algorithm that minimizes a layer-wise proxy loss. We theoretically prove this optimization is guaranteed to converge to a solution with a proxy loss less than or equal to state-of-the-art pruning algorithms. Experiments on Llama (Touvron et al., 2023; Dubey et al., 2024) and Qwen (Yang et al., 2025) model families demonstrate that ARMOR consistently and significantly outperforms state-of-the-art 2:4 pruning methods across a wide range of downstream tasks and perplexity evaluations. ARMOR achieves this superior performance while retaining the inference speedups and substantial memory usage reductions of 2:4 pruning, establishing a more effective trade-off between model compression and task accuracy

Core Task Landscape

This paper addresses: **Semi-Structured Pruning of Large Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

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

- **Semi-Structured Sparsity Pattern Methods**
- **One-Shot Importance-Based Pruning**
- **Depth and Layer-Level Pruning**
- **Hybrid Compression with Pruning**
- **Activation Sparsity and Inference Optimization**
- **Specialized Pruning Frameworks and Architectures**
- **Global and Coordinated Pruning Strategies**
- **Post-Training and Retraining-Free Methods**
- **Specialized Evaluation and Analysis**
- **Inference Acceleration and Hardware-Aware Pruning**
- ... and 1 more categories

Complete Taxonomy Tree

- Semi-Structured Pruning of Large Language Models Survey Taxonomy
- Semi-Structured Sparsity Pattern Methods
  - Learnable N:M Sparsity Mask Optimization ★ (5 papers)
  - [0] ARMOR: High-Performance Semi-Structured Pruning via Adaptive Matrix Factorization (Anon et al., 2026) View paper
  - [1] Maskllm: Learnable semi-structured sparsity for large language models (Fang, 2024) View paper
  - [36] CAST: Continuous and Differentiable Semi-Structured Sparsity-Aware Training for Large Language Models (Huang, 2025) View paper
  - [39] ProxSparse: Regularized Learning of Semi-Structured Sparsity Masks for Pretrained LLMs (Liu Hongyi, 2025) View paper
  - [40] Pruning large language models with semi-structural adaptive sparse training (Huang, 2025) View paper
  - Block-Wise and Structured Sparsity Patterns (5 papers)
  - [3] Blockpruner: Fine-grained pruning for large language models (Wan, 2025) View paper
  - [7] Pagedeviction: Structured block-wise kv cache pruning for efficient large language model inference (Chitty-Venkata, 2025) View paper
  - [27] From 2: 4 to 8: 16 sparsity patterns in LLMs for Outliers and Weights with Variance Correction (Egor Maximov, 2025) View paper
  - [30] Multipruner: Balanced structure removal in foundation models (MuÃ±oz, 2025) View paper
  - [46] Thanos: A Block-wise Pruning Algorithm for Efficient Large Language Model Compression (RichtÃ¡rik, 2025) View paper
  - Dependency-Aware and GLU-Specific Pruning (1 papers)
  - [38] Dependency-Aware Semi-Structured Sparsity of GLU Variants in Large Language Models (Guo Zhiyu, 2024) View paper
- One-Shot Importance-Based Pruning
  - Weight and Activation Magnitude Pruning (4 papers)

- [4] A simple and effective pruning approach for large language models (Sun MingJie, 2023) View paper
- [22] Wanda++: Pruning large language models via regional gradients (Yang, 2025) View paper
- [25] Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity (Yin Lu, 2023) View paper
- [44] NoWag: A Unified Framework for Shape Preserving Compression of Large Language Models (Liu, 2025) View paper
- Second-Order Information Pruning (3 papers)
- [10] The optimal bert surgeon: Scalable and accurate second-order pruning for large language models (Kurtic, 2022) View paper
- [23] Sparsegpt: Massive language models can be accurately pruned in one-shot (Frantar, 2023) View paper
- [33] BESA: Pruning large language models with blockwise parameter-efficient sparsity allocation (Xu Peng, 2024) View paper
- Entropy and Information-Theoretic Pruning (2 papers)
- [31] Entropy-Based Block Pruning for Efficient Large Language Models (Yang, 2025) View paper
- [42] LEP: Leveraging Local Entropy Pruning for Sparsity in Large Language Models (Yu-Li Chen, 2025) View paper

- Depth and Layer-Level Pruning
  - Static Depth Pruning (4 papers)
  - [5] Shortened llama: A simple depth pruning for large language models (Kim Bo Kyeong, 2024) View paper
  - [18] Shortened llama: Depth pruning for large language models with comparison of retraining methods (Kim Bo Kyeong, 2024) View paper
  - [32] Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks (Song Ji-won, 2024) View paper
  - [50] A deeper look at depth pruning of llms (Siddiqui, 2024) View paper
  - Dynamic Input-Aware Pruning (1 papers)
  - [47] IG-Pruning: Input-Guided Block Pruning for Large Language Models (Kangyu Qiao, 2025) View paper
- Hybrid Compression with Pruning
  - Joint Sparsity and Quantization (3 papers)
  - [14] Spqr: A sparse-quantized representation for near-lossless llm weight compression (Dettmers, 2023) View paper
  - [20] SPARQ: An accelerator architecture for large language models with joint sparsity and quantization techniques (Seonggyu Choi, 2025) View paper
  - [35] Compressing large language models by joint sparsification and quantization (J Guo, 2024) View paper
  - Low-Rank and Sparse Decomposition (3 papers)
  - [19] Losparse: Structured compression of large language models based on low-rank and sparse approximation (Li, 2023) View paper
  - [37] Large Language Model Compression with Global Rank and Sparsity Optimization (Zhou Chang-hai, 2025) View paper
  - [49] Efficient One-Shot Pruning of Large Language Models with Low-Rank Approximation (Yang-Yan Xu, 2024) View paper
  - Sparsity-Preserving Fine-Tuning (2 papers)
  - [11] Spp: Sparsity-preserved parameter-efficient fine-tuning for large language models (Lu Xudong, 2024) View paper
  - [43] SparseLoRA: Accelerating LLM Fine-Tuning with Contextual Sparsity (Khaki, 2025) View paper
- Activation Sparsity and Inference Optimization
  - Structured Activation Sparsity Learning (2 papers)
  - [17] Learn to be efficient: Build structured sparsity in large language models (Zheng Hai-zhong, 2024) View paper
  - [24] Novel Activation Sparsification Approach for Large Language Models (AV Demidovskij, 2025) View paper
  - Dynamic Activation Pruning (1 papers)
  - [29] R-sparse: Rank-aware activation sparsity for efficient llm inference (Zhang Zhen-yu, 2025) View paper
- Specialized Pruning Frameworks and Architectures
  - KV Cache and Memory-Efficient Pruning (1 papers)
  - [6] Kvpruner: Structural pruning for faster and memory-efficient large language models (Bo Lv, 2025) View paper
  - Sparse Attention Mechanisms (1 papers)
  - [41] The sparse frontier: Sparse attention trade-offs in transformer llms (Nawrot, 2025) View paper
  - State-Space Model Pruning (1 papers)
  - [45] SparseSSM: Efficient Selective Structured State Space Models Can Be Pruned in One-Shot (Wang Huan, 2025) View paper
- Global and Coordinated Pruning Strategies
  - Global Importance and Subproblem Decomposition (2 papers)
  - [2] Structured optimal brain pruning for large language models (Quan Lu, 2024) View paper
  - [16] SparseLLM: Towards global pruning of pre-trained language models (Guangji Bai, 2024) View paper
  - Collaborative Prompting and Structured Optimization (2 papers)
  - [15] Compresso: Structured pruning with collaborative prompting learns compact large language models (Guo, 2023) View paper
  - [34] NIRVANA: Structured pruning reimagined for large language models compression (Wei, 2025) View paper
- Post-Training and Retraining-Free Methods
  - Matrix Transformation and Slicing (1 papers)
  - [21] Slicegpt: Compress large language models by deleting rows and columns (Ashkboos, 2024) View paper
  - Layer-Wise Post-Training Pruning (2 papers)
  - [8] FISTAPruner: Layer-wise Post-training Pruning for Large Language Models (Pengxiang Zhao, 2025) View paper
  - [48] SlimLLM: Accurate Structured Pruning for Large Language Models (Guo Jia-long, 2025) View paper
- Specialized Evaluation and Analysis (2 papers)
  - [9] Model Hemorrhage and the Robustness Limits of Large Language Models (Ma, 2025) View paper
  - [12] Language-Specific Pruning for Efficient Reduction of Large Language Models (Shamrai, 2024) View paper
- Inference Acceleration and Hardware-Aware Pruning
  - GPU Kernel and SpMM Optimization (1 papers)
  - [13] Spinfer: Leveraging low-level sparsity for efficient large language model inference on gpus (Ruibo Fan, 2025) View paper
  - Sensitivity-Aware and Structured Compression (1 papers)
  - [28] Structured compression of large language models with sensitivity-aware pruning mechanisms (Yi-chen, 2024) View paper
- Task-Agnostic and Contrastive Pruning (1 papers)
  - [26] From dense to sparse: Contrastive pruning for better pre-trained language model compression (Runxin Xu, 2022) View paper

Narrative

Core task: semi-structured pruning of large language models. The field has organized itself around several complementary strategies for reducing model size and computational cost while preserving performance. At the highest level, Semi-Structured Sparsity Pattern Methods explore learnable or fixed N:M patterns that balance hardware efficiency with flexibility, while One-Shot Importance-Based Pruning techniques such as Sparsegpt[23] and Simple Effective Pruning[4] remove weights in a single pass using gradient or activation-based metrics. Depth and Layer-Level Pruning approaches like Shortened Llama[5] target entire layers or blocks, and Hybrid Compression with Pruning methods combine sparsity with quantization or low-rank factorization to achieve greater compression ratios. Meanwhile, Activation Sparsity and Inference Optimization branches focus on runtime efficiency, Specialized Pruning Frameworks adapt techniques to particular architectures, and Global and Coordinated Pruning Strategies enforce consistency across modules. Post-Training and Retraining-Free Methods aim to minimize calibration overhead, and dedicated evaluation branches assess the impact of pruning on downstream tasks and model behavior.

Within the Semi-Structured Sparsity Pattern Methods branch, a particularly active line of work centers on learnable N:M sparsity mask optimization, where methods such as Maskllm[1] and CAST[36] dynamically adjust which weights to retain during training or fine-tuning. ARMOR[0] sits squarely in this cluster, emphasizing adaptive mask learning to optimize the trade-off between sparsity ratio and task performance. Compared to ProxSparse[39], which employs proximal gradient techniques for mask updates, ARMOR[0] explores alternative optimization strategies that may offer faster convergence or better final accuracy. Similarly, Semi-Structural Adaptive[40] investigates adaptive sparsity patterns but differs in how it balances global versus local importance signals. Across these branches, key open questions include how to select optimal sparsity ratios without extensive retraining, how to coordinate pruning decisions across layers, and whether learnable masks can generalize across diverse tasks and model scales.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Maskllm: Learnable semi-structured sparsity for large language models

Authors: Fang, Gongfan, Yin, Hongxu, Gongfan Fang, et al. (23 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large Language Models (LLMs) are distinguished by their massive parameter counts, which typically result in significant redundancy. This work introduces MaskLLM, a learnable pruning method that establishes Semi-structured (or ``N:M'') Sparsity in LLMs, aimed at reducing computational overhead during inference. Instead of developing a new importance criterion, MaskLLM explicitly models N:M patterns as a learnable distribution through Gumbel Softmax sampling. This approach facilitates end-to-end t...

Relationship Analysis

Both papers belong to the Learnable N:M Sparsity Mask Optimization category, focusing on learning optimal 2:4 sparsity patterns for LLMs through gradient-based optimization rather than fixed importance criteria. While ARMOR uses a matrix factorization approach with block-diagonal wrappers around a sparse core optimized via block coordinate descent, MaskLLM models mask selection as a learnable categorical distribution using Gumbel Softmax sampling for end-to-end training. The key distinction is that ARMOR introduces additional low-overhead transformation matrices to provide flexibility while maintaining hardware acceleration, whereas MaskLLM directly learns mask probabilities through extensive iterative training on large-scale datasets.

---

2. CAST: Continuous and Differentiable Semi-Structured Sparsity-Aware Training for Large Language Models

Authors: Huang, Weiyu, Weiyu Huang, Zhu, Jun, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Sparsity-aware training is an effective approach for transforming large language models (LLMs) into hardware-friendly sparse patterns, thereby reducing latency and memory consumption during inference. In this paper, we propose Continuous Adaptive Sparse Trainer (CAST), a fully continuous and differentiable sparsity-aware training framework for semi-structured (or"N:M") sparse models. Unlike previous approaches that optimize sparsity patterns and weights separately, CAST enables seamless joint op...

Relationship Analysis

Both papers belong to the Learnable N:M Sparsity Mask Optimization category, focusing on gradient-based or regularized training to optimize semi-structured sparsity patterns. They overlap in addressing 2:4 sparsity for LLMs through learnable mask optimization during training, with both methods jointly optimizing masks and weights. However, ARMOR uses a one-shot post-training approach with matrix factorization (sparse core wrapped by block-diagonal matrices) and block coordinate descent, while CAST employs continuous sparsity-aware training from scratch or continued pretraining with an adaptive L1 decay optimizer (AdamS) and knowledge distillation, representing fundamentally different optimization paradigms (post-training factorization vs. training-time regularization).

---

3. ProxSparse: Regularized Learning of Semi-Structured Sparsity Masks for Pretrained LLMs

Authors: Liu Hongyi, Saha, Rajarshi, Hongyi Liu, Jia Zhen, et al. (20 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models (LLMs) have demonstrated exceptional performance in natural language processing tasks, yet their massive size makes serving them inefficient and costly. Semi-structured pruning has emerged as an effective method for model acceleration, but existing approaches are suboptimal because they focus on local, layer-wise optimizations using heuristic rules, failing to leverage global feedback. We present ProxSparse, a learning-based framework for mask selection enabled by regulariz...

Relationship Analysis

Both papers belong to the Learnable N:M Sparsity Mask Optimization category, focusing on learning-based approaches to discover optimal semi-structured sparsity patterns through gradient-based optimization. While ARMOR uses adaptive matrix factorization with block diagonal wrappers around a 2:4 sparse core optimized via block coordinate descent, ProxSparse employs regularized optimization with proximal gradient descent to learn masks directly through soft constraints that gradually enforce 2:4 sparsity. The key distinction is that ARMOR introduces additional low-rank structure through factorization to preserve model quality, whereas ProxSparse focuses on end-to-end mask learning with regularizers that transform rigid constraints into differentiable objectives without modifying the weight matrix structure.

---

4. Pruning large language models with semi-structural adaptive sparse training

Authors: Huang, Weiyu, Weiyu Huang, Guohao Jian, Zhu, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The remarkable success of Large Language Models (LLMs) relies heavily on their substantial scale, which poses significant challenges during model deployment in terms of latency and memory consumption. Recently, numerous studies have attempted to compress LLMs

using one-shot pruning methods. However, these methods often suffer from considerable performance degradation on complex language understanding tasks, raising concerns about the feasibility of pruning in LLMs. To address this issue, we prop...

Relationship Analysis

Both papers belong to the Learnable N:M Sparsity Mask Optimization category, focusing on methods that optimize sparsity masks through gradient-based training for semi-structured pruning of LLMs. While ARMOR uses a one-shot post-training approach with matrix factorization (block diagonal wrappers around a 2:4 sparse core) optimized via block coordinate descent, the candidate paper (AST) employs an adaptive sparse training framework that gradually learns optimal 2:4 masks during retraining with knowledge distillation and optional low-rank boosting parameters. The key distinction is that ARMOR is a post-training method without iterative retraining, whereas AST requires retraining with gradient-based mask updates and distillation from dense models.

Contributions Analysis

Overall novelty summary. The paper introduces ARMOR, a one-shot post-training pruning method that factorizes weight matrices into a 2:4 sparse core wrapped by block-diagonal error correctors. This work resides in the 'Learnable N:M Sparsity Mask Optimization' leaf, which contains five papers including the original submission. This leaf sits within the broader 'Semi-Structured Sparsity Pattern Methods' branch, indicating a moderately populated research direction focused on adaptive mask selection mechanisms. The taxonomy reveals that semi-structured pruning is an active area with multiple competing approaches across ten major branches.

The taxonomy tree shows that ARMOR's immediate neighbors include methods like MaskLLM and CAST, which also optimize N:M masks but through different training or fine-tuning strategies. Adjacent leaves explore 'Block-Wise and Structured Sparsity Patterns' (five papers) and 'Dependency-Aware and GLU-Specific Pruning' (one paper), suggesting that block-level transformations and architectural specialization are related but distinct research threads. The 'One-Shot Importance-Based Pruning' branch (nine papers across three leaves) represents an alternative paradigm using magnitude or Hessian-based metrics without learnable masks, highlighting a fundamental methodological divide in the field.

Among 29 candidates examined, none clearly refute any of ARMOR's three core contributions. The matrix factorization approach (10 candidates examined, 0 refutable) appears distinct from prior learnable mask methods in its use of block-diagonal wrappers rather than direct mask optimization. The block coordinate descent algorithm (9 candidates examined, 0 refutable) and convergence guarantee (10 candidates examined, 0 refutable) similarly show no substantial overlap within the limited search scope. These statistics suggest that ARMOR's combination of factorization, block-diagonal transformations, and theoretical guarantees may represent a novel synthesis, though the search examined only top-K semantic matches rather than an exhaustive literature review.

Based on the limited search scope of 29 candidates, ARMOR appears to occupy a relatively unexplored niche within learnable N:M sparsity optimization. The absence of refutable prior work across all three contributions, combined with its position in a moderately populated taxonomy leaf, suggests meaningful differentiation from existing approaches. However, this assessment is constrained by the top-K semantic search methodology and does not account for potentially relevant work outside the examined candidate set or in adjacent compression domains.

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

Contribution 1: ARMOR matrix factorization for semi-structured pruning

Description: The authors propose a novel weight representation that factorizes each weight matrix into a 2:4 sparse core surrounded by block diagonal wrapper matrices. These wrappers act as efficient pre- and post-transformation error correctors, offering greater flexibility to preserve model quality compared to conventional 2:4 pruning techniques.

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. Model compression and hardware acceleration for neural networks: A comprehensive survey

URL: View paper

Brief Assessment

Hardware Acceleration Survey[56] is a broad survey covering general compression techniques (compact models, tensor decomposition, quantization, network sparsification) without discussing the specific ARMOR factorization approach of using 2:4 sparse cores with block diagonal wrapper matrices as error correctors.

---

2. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification

URL: View paper

Brief Assessment

Singular Vector Orthogonality[54] focuses on low-rank matrix factorization via SVD training for general model compression, not specifically on semi-structured 2:4 sparsity patterns with block diagonal wrappers for hardware acceleration.

---

3. Sparse low rank factorization for deep neural network compression

URL: View paper

Brief Assessment

Sparse Low Rank[51] focuses on general low-rank factorization methods for neural network compression. The candidate does not address semi-structured 2:4 sparsity patterns or block diagonal wrapper matrices that are central to ARMOR's contribution.

---

4. A survey of deep neural network compression

URL: View paper

Brief Assessment

Compression Survey[57] only briefly mentions matrix decomposition methods in passing without providing technical details about factorization approaches for pruning with sparse cores and wrapper matrices.

---

5. TT@ CIM: A tensor-train in-memory-computing processor using bit-level-sparsity optimization and variable precision quantization

URL: View paper

Brief Assessment

Tensor-Train CIM[60] focuses on tensor-train decomposition for hardware-level CIM processors with bit-level sparsity optimization, not on matrix factorization methods for neural network weight pruning with sparse cores as in ARMOR.

---

6. Efficient neural network compression inspired by compressive sensing

URL: View paper

Brief Assessment

Compressive Sensing Compression[52] focuses on transform domain representations for general neural network compression, not on semi-structured 2:4 sparsity patterns with block diagonal wrappers for hardware acceleration in LLMs.

---

7. On compressing deep models by low rank and sparse decomposition

URL: View paper

Brief Assessment

Low Rank Sparse[55] focuses on decomposing weight matrices into low-rank and sparse components for general compression, not specifically for 2:4 semi-structured pruning with block diagonal wrappers as error correctors.

---

8. Group sparsity: The hinge between filter pruning and decomposition for network compression

URL: View paper

Brief Assessment

Group Sparsity[53] focuses on a unified framework for filter pruning and low-rank decomposition using group sparsity regularization on convolutional filters, not on semi-structured 2:4 pruning with block diagonal wrapper matrices for error correction as proposed in ARMOR.

---

9. Sparse convolutional neural networks

URL: View paper

Brief Assessment

Sparse Convolutional[59] focuses on sparse decomposition of convolutional kernels in CNNs using two-stage decompositions (channel basis and kernel basis), not the specific 2:4 semi-structured pruning with block diagonal wrapper matrices proposed in ARMOR for LLMs.

---

10. From galore to welore: How low-rank weights non-uniformly emerge from low-rank gradients

URL: View paper

Brief Assessment

Galore Welore[58] focuses on low-rank weight compression through gradient subspace stabilization analysis, not semi-structured 2:4 pruning with sparse cores and block diagonal wrappers as proposed in ARMOR.

Contribution 2: Block coordinate descent optimization algorithm

Description: The authors develop a block coordinate descent optimization algorithm that alternates between updating continuous parameters (the block diagonal matrices and dense weights) and updating the sparse core. This algorithm is designed to minimize a layer-wise proxy loss while respecting the 2:4 sparsity constraint.

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. Block coordinate descent algorithms for large-scale sparse multiclass classification

URL: View paper

Brief Assessment

Block Coordinate Descent[63] addresses sparse multiclass classification with $\ell 1/\ell 2$ regularization, not semi-structured pruning of neural networks with 2:4 sparsity constraints and matrix factorization wrappers.

---

2. A block decomposition algorithm for sparse optimization

URL: View paper

Brief Assessment

Block Decomposition Algorithm[67] focuses on sparse optimization problems with $\ell 0$ constraints, using combinatorial search over small working sets. The ORIGINAL paper addresses semi-structured 2:4 sparsity in neural network pruning with matrix factorization, a fundamentally different problem domain and approach.

---

3. Design and application of adaptive sparse deep echo state network

URL: View paper

Brief Assessment

Adaptive Sparse Echo[66] uses coordinate descent for sparse output weight learning in echo state networks for time series forecasting, not for optimizing sparse neural network weights with 2:4 sparsity constraints in LLMs.

---

4. Algorithmic and theoretical aspects of sparse deep neural networks

URL: View paper

Brief Assessment

Algorithmic Theoretical Aspects[61] focuses on sparse matrix factorization problems and their theoretical properties (ill-posedness, NP-hardness, optimization landscape). The thesis does not address block coordinate descent algorithms for optimizing sparse neural network weights with 2:4 sparsity constraints or layer-wise proxy loss minimization as described in the original paper.

---

5. Efficient blind source separation method for fMRI using autoencoder and spatiotemporal sparsity constraints

URL: View paper

Brief Assessment

Blind Source Separation[68] applies block coordinate descent to fMRI signal processing with sparsity constraints, not to neural network weight pruning or matrix factorization for model compression. The application domains and optimization objectives are fundamentally different.

---

6. Training block-wise sparse models using kronecker product decomposition

URL: View paper

Brief Assessment

Kronecker Product Decomposition[64] focuses on training block-wise sparse models using Kronecker product factorization with gradient descent updates, not on alternating block coordinate descent for 2:4 sparsity patterns with sparse core and block diagonal wrappers as in the original paper.

---

7. SequentialAttention++ for Block Sparsification: Differentiable Pruning Meets Combinatorial Optimization

URL: View paper

Brief Assessment

SequentialAttention++[69] focuses on combining differentiable pruning with combinatorial optimization for block sparsification, using local search methods inspired by iterative hard thresholding. The ORIGINAL paper's block coordinate descent algorithm alternates between updating continuous parameters (block diagonal matrices and dense weights) and updating a 2:4 sparse core with specific hardware constraints, which is a fundamentally different optimization approach targeting semi-structured sparsity patterns.

---

8. STICKER-IM: A 65 nm computing-in-memory NN processor using block-wise sparsity optimization and inter/intra-macro data reuse

URL: View paper

Brief Assessment

STICKER-IM[65] focuses on hardware architecture for computing-in-memory neural network processors with block-wise sparsity optimization, not on developing block coordinate descent algorithms for weight optimization.

---

9. BESA: Pruning large language models with blockwise parameter-efficient sparsity allocation

URL: View paper

Brief Assessment

BESA[33] uses gradient descent to optimize sparsity allocation parameters, not block coordinate descent for alternating between continuous parameters and sparse core updates as in the original paper.

Contribution 3: Theoretical convergence guarantee

Description: The authors establish a theoretical result (Theorem 3.1) proving that their optimization algorithm converges and achieves a proxy loss no worse than state-of-the-art methods like NoWag-P, providing formal guarantees for their approach.

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. Compression aware training of neural networks using Frank-Wolfe

URL: View paper

Brief Assessment

Frank-Wolfe Compression[70] addresses convergence of the Frank-Wolfe algorithm for neural network compression, not pruning optimization algorithms. The candidate's Theorem 4.1 proves convergence for gradient rescaling in SFW, while the original paper's Theorem 3.1 proves convergence for a block coordinate descent algorithm with specific proxy loss guarantees for pruning.

---

2. Concurrent training and layer pruning of deep neural networks

URL: View paper

Brief Assessment

Concurrent Training Pruning[78] focuses on layer pruning with convergence guarantees for their specific optimization algorithm involving Bernoulli random variables and projected gradient descent. The ORIGINAL paper addresses semi-structured 2:4 pruning with matrix factorization and block coordinate descent, representing fundamentally different pruning approaches and optimization frameworks.

---

3. The convergence of sparsified gradient methods

URL: View paper

Brief Assessment

Sparsified Gradient Convergence[72] analyzes gradient sparsification methods for distributed training, not neural network pruning optimization. The candidate focuses on communication reduction in distributed SGD by selecting top-k gradient components, while the original paper addresses post-training pruning with matrix factorization for model compression.

---

4. Directional pruning of deep neural networks

URL: View paper

Brief Assessment

Directional Pruning[74] provides convergence guarantees for a directional pruning algorithm based on GRDA, not for semi-structured 2:4 pruning with matrix factorization. The theoretical frameworks address fundamentally different optimization problems and pruning approaches.

---

5. Mask in the mirror: Implicit sparsification

URL: View paper

Brief Assessment

Mask Mirror[75] focuses on continuous sparsification with implicit bias and mirror flow convergence for pruning optimization, not the specific 2:4 semi-structured pruning with matrix factorization that ARMOR addresses. The theoretical frameworks and problem formulations differ fundamentally.

---

6. Fast convex pruning of deep neural networks

URL: View paper

Brief Assessment

Fast Convex Pruning[73] focuses on layer-wise convex optimization with sample complexity guarantees for sparse weight discovery, not on convergence guarantees for iterative pruning optimization algorithms like ARMOR's block coordinate descent method.

---

7. Enhancing personalized model construction and privacy protection in federated learning with generative adversarial networks and parameter sparsification

URL: View paper

Brief Assessment

Federated GAN Sparsification[71] focuses on federated learning with GANs and parameter sparsification for privacy protection, not neural network pruning optimization algorithms. The minimal context provided does not contain sufficient technical detail about convergence guarantees for pruning methods to assess overlap with ARMOR's Theorem 3.1.

---

8. Efficient construction and convergence analysis of sparse convolutional neural networks

URL: View paper

Brief Assessment

Sparse CNN Construction[77] focuses on sparse convolutional neural network construction methods and does not address convergence guarantees for pruning optimization algorithms in the context of large language models or semi-structured sparsity patterns.

---

9. Net-trim: Convex pruning of deep neural networks with performance guarantee

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

10. Optimal approximation with sparsely connected deep neural networks

URL: View paper

Brief Assessment

Optimal Approximation Sparse[76] focuses on approximation theory for neural networks and function classes, establishing lower bounds on connectivity and memory for achieving uniform approximation rates. This is fundamentally different from ARMOR's convergence guarantee for a pruning optimization algorithm that minimizes a proxy loss.

Appendix: Text Similarity Detection

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

References

- [0] ARMOR: High-Performance Semi-Structured Pruning via Adaptive Matrix Factorization View paper
- [1] Maskllm: Learnable semi-structured sparsity for large language models View paper
- [2] Structured optimal brain pruning for large language models View paper
- [3] Blockpruner: Fine-grained pruning for large language models View paper
- [4] A simple and effective pruning approach for large language models View paper
- [5] Shortened llama: A simple depth pruning for large language models View paper
- [6] Kvpruner: Structural pruning for faster and memory-efficient large language models View paper
- [7] Pagedeviction: Structured block-wise kv cache pruning for efficient large language model inference View paper
- [8] FISTAPruner: Layer-wise Post-training Pruning for Large Language Models View paper
- [9] Model Hemorrhage and the Robustness Limits of Large Language Models View paper
- [10] The optimal bert surgeon: Scalable and accurate second-order pruning for large language models View paper
- [11] Spp: Sparsity-preserved parameter-efficient fine-tuning for large language models View paper
- [12] Language-Specific Pruning for Efficient Reduction of Large Language Models View paper
- [13] Spinfer: Leveraging low-level sparsity for efficient large language model inference on gpus View paper
- [14] Spqr: A sparse-quantized representation for near-lossless llm weight compression View paper
- [15] Compresso: Structured pruning with collaborative prompting learns compact large language models View paper
- [16] SparseLLM: Towards global pruning of pre-trained language models View paper
- [17] Learn to be efficient: Build structured sparsity in large language models View paper
- [18] Shortened llama: Depth pruning for large language models with comparison of retraining methods View paper
- [19] Losparse: Structured compression of large language models based on low-rank and sparse approximation View paper
- [20] SPARQ: An accelerator architecture for large language models with joint sparsity and quantization techniques View paper
- [21] Slicegpt: Compress large language models by deleting rows and columns View paper
- [22] Wanda++: Pruning large language models via regional gradients View paper
- [23] Sparsegpt: Massive language models can be accurately pruned in one-shot View paper
- [24] Novel Activation Sparsification Approach for Large Language Models View paper
- [25] Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity View paper
- [26] From dense to sparse: Contrastive pruning for better pre-trained language model compression View paper
- [27] From 2: 4 to 8: 16 sparsity patterns in LLMs for Outliers and Weights with Variance Correction View paper
- [28] Structured compression of large language models with sensitivity-aware pruning mechanisms View paper
- [29] R-sparse: Rank-aware activation sparsity for efficient llm inference View paper
- [30] Multipruner: Balanced structure removal in foundation models View paper
- [31] Entropy-Based Block Pruning for Efficient Large Language Models View paper
- [32] Sleb: Streamlining llms through redundancy verification and elimination of transformer blocks View paper
- [33] BESA: Pruning large language models with blockwise parameter-efficient sparsity allocation View paper
- [34] NIRVANA: Structured pruning reimagined for large language models compression View paper
- [35] Compressing large language models by joint sparsification and quantization View paper
- [36] CAST: Continuous and Differentiable Semi-Structured Sparsity-Aware Training for Large Language Models View paper
- [37] Large Language Model Compression with Global Rank and Sparsity Optimization View paper
- [38] Dependency-Aware Semi-Structured Sparsity of GLU Variants in Large Language Models View paper
- [39] ProxSparse: Regularized Learning of Semi-Structured Sparsity Masks for Pretrained LLMs View paper
- [40] Pruning large language models with semi-structural adaptive sparse training View paper
- [41] The sparse frontier: Sparse attention trade-offs in transformer llms View paper
- [42] LEP: Leveraging Local Entropy Pruning for Sparsity in Large Language Models View paper
- [43] SparseLoRA: Accelerating LLM Fine-Tuning with Contextual Sparsity View paper
- [44] NoWag: A Unified Framework for Shape Preserving Compression of Large Language Models View paper
- [45] SparseSSM: Efficient Selective Structured State Space Models Can Be Pruned in One-Shot View paper
- [46] Thanos: A Block-wise Pruning Algorithm for Efficient Large Language Model Compression View paper
- [47] IG-Pruning: Input-Guided Block Pruning for Large Language Models View paper
- [48] SlimLLM: Accurate Structured Pruning for Large Language Models View paper

- [49] Efficient One-Shot Pruning of Large Language Models with Low-Rank Approximation View paper
- [50] A deeper look at depth pruning of llms View paper
- [51] Sparse low rank factorization for deep neural network compression View paper
- [52] Efficient neural network compression inspired by compressive sensing View paper
- [53] Group sparsity: The hinge between filter pruning and decomposition for network compression View paper
- [54] Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification View paper
- [55] On compressing deep models by low rank and sparse decomposition View paper
- [56] Model compression and hardware acceleration for neural networks: A comprehensive survey View paper
- [57] A survey of deep neural network compression View paper
- [58] From galore to welore: How low-rank weights non-uniformly emerge from low-rank gradients View paper
- [59] Sparse convolutional neural networks View paper
- [60] TT@ CIM: A tensor-train in-memory-computing processor using bit-level-sparsity optimization and variable precision quantization View paper
- [61] Algorithmic and theoretical aspects of sparse deep neural networks View paper
- [62] Ebft: Effective and block-wise fine-tuning for sparse llms View paper
- [63] Block coordinate descent algorithms for large-scale sparse multiclass classification View paper
- [64] Training block-wise sparse models using kronecker product decomposition View paper
- [65] STICKER-IM: A 65 nm computing-in-memory NN processor using block-wise sparsity optimization and inter/intra-macro data reuse View paper
- [66] Design and application of adaptive sparse deep echo state network View paper
- [67] A block decomposition algorithm for sparse optimization View paper
- [68] Efficient blind source separation method for fMRI using autoencoder and spatiotemporal sparsity constraints View paper
- [69] SequentialAttention++ for Block Sparsification: Differentiable Pruning Meets Combinatorial Optimization View paper
- [70] Compression aware training of neural networks using Frank-Wolfe View paper
- [71] Enhancing personalized model construction and privacy protection in federated learning with generative adversarial networks and parameter sparsification View paper
- [72] The convergence of sparsified gradient methods View paper
- [73] Fast convex pruning of deep neural networks View paper
- [74] Directional pruning of deep neural networks View paper
- [75] Mask in the mirror: Implicit sparsification View paper
- [76] Optimal approximation with sparsely connected deep neural networks View paper
- [77] Efficient construction and convergence analysis of sparse convolutional neural networks View paper
- [78] Concurrent training and layer pruning of deep neural networks View paper
- [79] Net-trim: Convex pruning of deep neural networks with performance guarantee View paper