Novelty Assessment Report

Paper: MoNE: Replacing Redundant Experts with Lightweight Novices for Structured Pruning of MoE
PDF URL: https://openreview.net/pdf?id=881uEwToKQ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-08

Abstract

Mixture-of-Experts (MoE) enables efficient scaling of large language models by activating only a subset of experts per input token. However, deploying MoE-based models incurs significant memory overhead due to the need to retain all experts in memory. While structured pruning is promising to reduce memory costs, existing methods often show suboptimal performance and unstable degradation in three dimensions: model architectures, calibration data sources, and calibration sample sizes. This paper proposes \textbf{M}ixture-\textbf{o}f-\textbf{N}ovices-and-\textbf{E}xperts (\textbf{MoNE}), a novel expert pruning method that replaces redundant experts with lightweight novices to achieve effective and robust model compression. MoNE evaluates expert redundancy based on two metrics: access frequency and output variance. Experts exhibiting low usage and stable outputs are pruned and replaced with lightweight novices—unbiased estimations of their original outputs—minimizing performance degradation. Extensive experiments demonstrate that MoNE consistently outperforms baseline methods with minimal accuracy degradation across the three dimensions, confirming its effectiveness and robustness. Notably, it outperforms baselines by up to 2.72 for the average zero shot accuracy across nine downstream tasks under 25\% pruning ratio, with only 0.14 performance drop for Qwen2-57B-A14B. The code is available at https://anonymous.4open.science/r/AnonymizedMoNE.

Core Task Landscape

This paper addresses: **structured pruning of mixture-of-experts 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:

- **Expert-Level Structured Pruning Methods**
- **Intra-Expert Compression Techniques**
- **Hybrid Compression Frameworks**
- **Structured Sparsity and Activation Optimization**
- **Theoretical Foundations and Analysis**
- **Adaptive and Learnable Compression**
- **One-Shot Pruning Strategies**
- **MoE Construction and Conversion**
- **Surveys and Systematic Reviews**
- **Specialized MoE Applications and Architectures**
- ... and 4 more categories

Complete Taxonomy Tree

- structured pruning of mixture-of-experts models Survey Taxonomy
- Expert-Level Structured Pruning Methods
  - Redundancy-Based Expert Removal
  - Clustering-Driven Expert Pruning (2 papers)
    - [1] Cluster-Driven Expert Pruning for Mixture-of-Experts Large Language Models (Guo, 2025) View paper
    - [19] Diversifying the Expert Knowledge for Task-Agnostic Pruning in Sparse Mixture-of-Experts (Zhang, 2024) View paper
  - Direct Similarity-Based Pruning (2 papers)
    - [24] Mosaic Pruning: A Hierarchical Framework for Generalizable Pruning of Mixture-of-Experts Models (Wentao Hu, 2025) View paper
    - [33] Finding Fantastic Experts in MoEs: A Unified Study for Expert Dropping Strategies and Observations (Jaiswal, 2025) View paper
  - Usage-Guided Expert Selection
  - Activation Frequency and Routing Analysis (3 papers)
    - [6] Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models (Huang, 2024) View paper
    - [22] Seer-moe: Sparse expert efficiency through regularization for mixture-of-experts (Muzio, 2024) View paper
    - [36] Moe-pruner: Pruning mixture-of-experts large language model using the hints from its router (Xie, 2024) View paper
  - Output Stability-Based Pruning ★ (2 papers)
    - [0] MoNE: Replacing Redundant Experts with Lightweight Novices for Structured Pruning of MoE (Anon et al., 2026) View paper
    - [5] Mixture Compressor for Mixture-of-Experts LLMs Gains More (Huang Wei, 2024) View paper
  - Task-Specific and Domain-Adaptive Pruning (3 papers)
  - [3] Domain-Specific Pruning of Large Mixture-of-Experts Models with Few-shot Demonstrations (Dong, 2025) View paper

- [18] Task-Specific Expert Pruning for Sparse Mixture-of-Experts (Chen Tianyu, 2022) View paper
- [20] A Provably Effective Method for Pruning Experts in Fine-tuned Sparse Mixture-of-Experts (Chowdhury, 2024) View paper

- Intra-Expert Compression Techniques
  - Quantization-Based Expert Compression (2 papers)
  - [2] EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models (Chen Yuan-Teng, 2025) View paper
  - [14] MC#: Mixture Compressor for Mixture-of-Experts Large Models (Huang Wei, 2025) View paper
  - Low-Rank Decomposition and Weight Sharing (2 papers)
  - [12] Delta Decompression for MoE-based LLMs Compression (Gu Hao, 2025) View paper
  - [21] MoE-IÂ²: Compressing Mixture of Experts Models through Inter-Expert Pruning and Intra-Expert Low-Rank Decomposition (Cheng Yang, 2024) View paper
- Hybrid Compression Frameworks
  - Pruning with Distillation or Fine-Tuning (1 papers)
  - [8] SlimMoE: Structured Compression of Large MoE Models via Expert Slimming and Distillation (Li, 2025) View paper
  - Multi-Stage Compression Pipelines (1 papers)
  - [35] Towards Efficient Mixture of Experts: A Holistic Study of Compression Techniques (He, 2024) View paper
  - Expert Merging and Recombination (4 papers)
  - [25] REAP the Experts: Why Pruning Prevails for One-Shot MoE compression (Lasby, 2025) View paper
  - [32] Dropping Experts, Recombining Neurons: Retraining-Free Pruning for Sparse Mixture-of-Experts LLMs (Yixiao Zhou, 2025) View paper
  - [40] Condense, Don't Just Prune: Enhancing Efficiency and Performance in MoE Layer Pruning (Cao Ming-yu, 2024) View paper
  - [41] PuzzleMoE: Efficient Compression of Large Mixture-of-Experts Models via Sparse Expert Merging and Bit-packed inference (Zhao Yushu, 2025) View paper
- Structured Sparsity and Activation Optimization
  - Hardware-Aware Structured Sparsity (2 papers)
  - [9] Samoyeds: Accelerating MoE Models with Structured Sparsity Leveraging Sparse Tensor Cores (Wu Chenpeng, 2025) View paper
  - [38] Flame: Fully leveraging moe sparsity for transformer on fpga (Xuanda Lin, 2024) View paper
  - Dynamic Routing and Gating Sparsification (4 papers)
  - [13] Dense-to-sparse gate for mixture-of-experts (X Nie, 2021) View paper
  - [23] Dynamic Mixture of Experts for Adaptive Computation in Character-Level Transformers (Zhi-gao, 2025) View paper
  - [30] Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation (Dong, 2024) View paper
  - [45] Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate (Nie, 2021) View paper
- Theoretical Foundations and Analysis
  - Empirical Analysis of Sparsity (3 papers)
  - [34] Llama-moe v2: Exploring sparsity of llama from perspective of mixture-of-experts with post-training (Qu, 2024) View paper
  - [39] Sparse Mixture-of-Experts for Compositional Generalization: Empirical Evidence and Theoretical Foundations of Optimal Sparsity (Zhao, 2024) View paper
  - [50] Mixture of Neuron Experts (Cheng RunXi, 2025) View paper
- Adaptive and Learnable Compression
  - Differentiable Pruning Optimization (1 papers)
  - [49] DiEP: Adaptive Mixture-of-Experts Compression through Differentiable Expert Pruning (Bai, 2025) View paper
- One-Shot Pruning Strategies (1 papers)
  - [16] STUN: Structured-Then-Unstructured Pruning for Scalable MoE Pruning (Campos, 2024) View paper
- MoE Construction and Conversion (2 papers)
  - [10] ToMoE: Converting Dense Large Language Models to Mixture-of-Experts through Dynamic Structural Pruning (Gao, 2025) View paper
  - [43] Dense2MoE: Restructuring Diffusion Transformer to MoE for Efficient Text-to-Image Generation (Zheng You-wei, 2025) View paper
- Surveys and Systematic Reviews (1 papers)
  - [4] A survey on inference optimization techniques for mixture of experts models (Liu Jiaâ□□Cheng, 2024) View paper
- Specialized MoE Applications and Architectures
  - Vision and Multimodal MoE (4 papers)
  - [26] Edge-moe: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts (Rishov Sarkar, 2023) View paper
  - [27] Sparse-MoE-SAM: A Lightweight Framework Integrating MoE and SAM with a Sparse Attention Mechanism for Plant Disease Segmentation in Resource â□¦ (B Zhao, 2025) View paper
  - [31] Scaling Vision with Sparse Mixture of Experts (Carlos Riquelme, 2022) View paper
  - [47] Efficient Multimodal Streaming Recommendation via Expandable Side Mixture-of-Experts (Yunke Qu, 2025) View paper
  - Continual and Lifelong Learning MoE (1 papers)
  - [29] R^ 2MoE: Redundancy-Removal Mixture of Experts for Lifelong Concept Learning (Guo, 2025) View paper
  - Domain-Specific MoE Architectures (2 papers)
  - [15] Mixture of Length and Pruning Experts for Knowledge Graphs Reasoning (Liu Siyi, 2025) View paper
  - [37] Demix layers: Disentangling domains for modular language modeling (S Gururangan, 2022) View paper
- Alternative Sparsity Mechanisms (3 papers)
  - [17] From Sparse to Soft Mixtures of Experts (Puigcerver, 2023) View paper
  - [46] ReMod: Learning Structured Sparsity with ReLU Modulation (W Zhang, 2025) View paper
  - [48] Sparse mixers: Combining moe and mixing to build a more efficient bert (Lee-Thorp, 2022) View paper
- Foundational MoE Concepts (3 papers)
  - [7] Mixture of experts (moe): A big data perspective (Wensheng Gan, 2025) View paper
  - [42] Structural adaptation in mixture of experts (V. Ramamurti, 1996) View paper
  - [44] StructPrune: Structured Global Pruning asymptotics with GPU Memory (X Song, 2025) View paper
- Pre-Training Efficiency (1 papers)
  - [28] Efficient Expert Pruning for Pre-Training of Mixture-of-Experts Large Language Models (S Wu, 2025) View paper

- Comparative Pruning Studies (1 papers)
  - [11] Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models: Enhancing Performance and Reducing Inference Costs (Liu, 2024) View paper

Narrative

Core task: structured pruning of mixture-of-experts models. The field organizes around several complementary strategies for compressing MoE architectures while preserving their conditional computation benefits. Expert-Level Structured Pruning Methods focus on removing or merging entire experts based on usage patterns, output stability, or clustering criteria, as seen in works like Cluster Expert Pruning[1] and Not All Experts[6]. Intra-Expert Compression Techniques instead target redundancy within individual experts through weight pruning or low-rank decomposition, while Hybrid Compression Frameworks combine expert-level and intra-expert approaches for greater efficiency gains. Additional branches address Adaptive and Learnable Compression (where pruning decisions evolve during training), One-Shot Pruning Strategies (enabling rapid post-training compression), and MoE Construction and Conversion (transforming dense models into sparse MoE variants). Theoretical Foundations and Analysis provide formal guarantees, and Surveys and Systematic Reviews like MoE Inference Survey[4] synthesize emerging trends across these diverse methodologies.

A particularly active line of work explores usage-guided expert selection, where pruning decisions rely on tracking which experts contribute meaningfully across different inputs or domains. MoNE[0] exemplifies this direction by emphasizing output stability-based pruning, ensuring that removed experts minimally disrupt model predictions. This approach contrasts with simpler frequency-based methods and aligns closely with Mixture Compressor[5], which also prioritizes preserving critical expert contributions during compression. Meanwhile, domain-specific strategies like Domain Specific Pruning[3] tailor expert removal to particular task distributions, and clustering-based methods such as Cluster Expert Pruning[1] group redundant experts before merging. MoNE[0] sits within the usage-guided cluster, distinguished by its focus on output stability rather than purely activation frequency, offering a middle ground between aggressive one-shot pruning and computationally intensive adaptive retraining schemes.

Related Works in Same Category

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

1. Mixture Compressor for Mixture-of-Experts LLMs Gains More

Authors: Huang Wei, Liao, Yue, Liu Jian-hui, He, et al. (13 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Mixture-of-Experts large language models (MoE-LLMs) marks a significant step forward of language models, however, they encounter two critical challenges in practice: 1) expert parameters lead to considerable memory consumption and loading latency; and 2) the current activated experts are redundant, as many tokens may only require a single expert. Motivated by these issues, we investigate the MoE-LLMs and make two key observations: a) different experts exhibit varying behaviors on activation reco...

Relationship Analysis

Both papers belong to the Output Stability-Based Pruning category, focusing on identifying and removing experts with stable or low-variance outputs in MoE models. The original paper (MoNE) evaluates expert redundancy using both access frequency and output variance, then replaces pruned experts with lightweight novices (unbiased output estimations), while the candidate paper (Mixture Compressor) combines static mixed-precision quantization with dynamic pruning, using activation reconstruction error, routing scores, and frequencies to determine expert importance. The key difference is that MoNE focuses on expert replacement with constant vectors to minimize output discrepancy, whereas Mixture Compressor integrates quantization and dynamic pruning without explicit expert replacement mechanisms.

Contributions Analysis

Overall novelty summary. The paper proposes MoNE, a framework that replaces redundant experts with lightweight 'novices' to compress MoE models. It resides in the 'Output Stability-Based Pruning' leaf under 'Usage-Guided Expert Selection', which contains only two papers including this one. This leaf is part of the broader 'Expert-Level Structured Pruning Methods' branch, which encompasses multiple pruning strategies across six leaves. The sparse population of this specific leaf suggests that output stability as a primary pruning criterion remains relatively underexplored compared to frequency-based or clustering-driven approaches.

The taxonomy reveals that MoNE's immediate neighbors include 'Activation Frequency and Routing Analysis' (three papers) and 'Redundancy-Based Expert Removal' (five papers across two sub-leaves). The broader 'Expert-Level Structured Pruning Methods' branch contains eleven papers total, while sibling branches like 'Intra-Expert Compression Techniques' and 'Hybrid Compression Frameworks' offer complementary compression strategies. MoNE's focus on output variance distinguishes it from frequency-only methods in the neighboring leaf, yet it shares conceptual overlap with redundancy detection approaches that measure expert similarity through output comparisons rather than weight-space metrics.

Among twenty-two candidates examined, the contribution-level analysis shows mixed novelty signals. The core MoNE framework (Contribution A) examined ten candidates and found one potentially refuting prior work, suggesting some overlap exists within the limited search scope. The dual-metric redundancy evaluation using access frequency and output variance (Contribution B) examined two candidates with no clear refutations, indicating this specific combination may be less explored. The robustness claim across architectures, data sources, and sample sizes (Contribution C) examined ten candidates without refutation, though this may reflect the limited scope rather than definitive novelty.

Based on the top-22 semantic matches examined, MoNE appears to occupy a moderately explored niche within usage-guided pruning. The sparse leaf population and limited refutations for two of three contributions suggest potential novelty, though the single refutation for the core framework warrants careful examination of overlapping prior work. The analysis does not cover exhaustive citation networks or domain-specific venues, leaving open questions about related work in specialized MoE compression literature.

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

Contribution 1: MoNE: Mixture-of-Novices-and-Experts framework

Description: MoNE is a novel expert pruning method for compressing MoE models by replacing redundant experts with lightweight novices (unbiased estimations of expert outputs) rather than simply removing them, thereby minimizing performance degradation while reducing memory overhead.

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. Unveiling super experts in mixture-of-experts large language models

URL: View paper

Brief Assessment

Super Experts[53] focuses on identifying and analyzing a small subset of critical experts (super experts) that are essential for model performance, rather than proposing a compression method that replaces redundant experts with lightweight structures. The papers

address different aspects of MoE models: MoNE proposes a novel pruning framework with novice replacement, while Super Experts[53] characterizes the importance and mechanisms of specific critical experts.

---

2. Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models: Enhancing Performance and Reducing Inference Costs

URL: View paper

Brief Assessment

Efficient Expert Pruning[11] focuses on evolutionary search-based expert pruning and merging for MoE models, while MoNE introduces a novel framework that replaces pruned experts with lightweight novices (unbiased estimations). These are fundamentally different compression approaches with distinct technical mechanisms.

---

3. Cluster-Driven Expert Pruning for Mixture-of-Experts Large Language Models

URL: View paper

Brief Assessment

Cluster Expert Pruning[1] focuses on clustering functionally similar experts and eliminating redundant clusters, while MoNE replaces redundant experts with lightweight novices (unbiased estimations). These are fundamentally different compression strategies - clustering-based elimination versus novice replacement.

---

4. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models

URL: View paper

Brief Assessment

MoE LLaVA[55] focuses on vision-language models with mixture-of-experts for multi-modal understanding, not on expert pruning methods for compressing MoE models. The candidate addresses model architecture design for LVLMs, while the original contribution is about structured pruning and compression techniques.

---

5. A survey on inference optimization techniques for mixture of experts models

URL: View paper

Brief Assessment

MoE Inference Survey[4] is a comprehensive survey paper that reviews existing optimization techniques for MoE models, including expert pruning methods. It does not present a novel expert pruning method that would refute MoNE's novelty claim of replacing redundant experts with lightweight novices.

---

6. Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models

URL: View paper

Prior Art Analysis

Not All Experts[6] demonstrates that expert pruning methods for MoE models existed prior to the original paper's submission. Both papers address the same core problem: reducing memory overhead in MoE models by removing redundant experts while minimizing performance degradation. Not All Experts[6] proposes post-training expert pruning methods that permanently discard less important experts based on reconstruction loss minimization, achieving similar goals of memory reduction and performance preservation. The candidate paper's approach of enumerating expert combinations and selecting those with lowest reconstruction loss directly challenges the novelty of replacing experts with lightweight structures, as it shows that simply removing experts (without replacement) can achieve effective compression.

Evidence

Evidence 1 - **Rationale**: Not All Experts[6] proposes post-training expert pruning methods that permanently remove experts to reduce parameters and improve efficiency, demonstrating that expert pruning approaches existed before MoNE's novelty claim of replacing experts with novices. - **Original**: this paper proposesmixture-of-novices-and-experts (mone), a novel expert pruning method that replaces redundant experts with lightweight novices to achieve effective and robust model compression. - **Candidate**: we introduce a heuristic search method to prune the number of experts in a post-training manner. task-agnostic expert pruning for general tasks. different from existing pruning schemes leveraging unstructured or semi-structured weight sparsity (sun et al., 2023; frantar and alistarh, 2023) for llms,...

Evidence 2 - **Rationale**: Not All Experts[6] shows that expert pruning based on reconstruction loss minimization was already proposed, which directly relates to the core mechanism MoNE uses to identify redundant experts, though MoNE adds the novice replacement step. - **Original**: to improve the effectiveness and robustness of structured pruning for moe models, this paper proposes a novel expert pruning method,mixture-of-novices-and-experts (mone) which replaces redundant experts with a lightweight structure,novice. - **Candidate**: we meticulously enumerate and choose combinations of experts that yield the lowest token reconstruction loss, subsequently, concatenating them to obtain the final pruned moe model. this strategy significantly lowers the memory demands for deploying moe llms.

Evidence 3 - **Rationale**: Not All Experts[6] demonstrates that evaluating expert importance through reconstruction loss (which implicitly captures output stability) was already established, challenging the novelty of MoNE's variance-based redundancy metric as a fundamentally new approach. - **Original**: mone evaluates expert redundancy based on two metrics: access frequency and output variance. experts exhibiting low usage and stable outputs are pruned and replaced with lightweight novices-unbiased estimations of their original outputs-minimizing performance degradation. - **Candidate**: the frobenius norm of the difference between cached output f(x) and the output of pruned layer f′(x, c) is used to quantify the discrepancy between the model before and after expert pruning, and we denote it as reconstruction loss. the expert subset corresponding to the minimum reconstruction loss i...

---

7. Mixture Compressor for Mixture-of-Experts LLMs Gains More

URL: View paper

Brief Assessment

Mixture Compressor[5] focuses on quantization and dynamic pruning for MoE compression, not on replacing redundant experts with lightweight novices as MoNE does. The candidate uses mixed-precision quantization and online dynamic pruning, which are fundamentally different compression techniques.

---

8. Task-Specific Expert Pruning for Sparse Mixture-of-Experts

URL: View paper

Brief Assessment

Task Specific Pruning[18] focuses on progressively dropping non-professional experts during fine-tuning for downstream tasks, converting MoE to single-expert dense models. MoNE addresses a different problem: compressing MoE models by replacing redundant

experts with lightweight novices (constant vectors) rather than removing them entirely, targeting deployment memory overhead rather than task-specific adaptation.

---

9. A closer look into mixture-of-experts in large language models

URL: View paper

Brief Assessment

Closer Look MoE[54] focuses on analyzing existing MoE architectures through parameter and behavioral studies, not on proposing expert pruning methods for model compression. The paper investigates expert similarities, routing mechanisms, and output correlations to understand MoE inner workings, which is fundamentally different from MoNE's contribution of replacing redundant experts with lightweight novices for compression.

---

10. EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models

URL: View paper

Brief Assessment

EAC-MoE[2] focuses on quantization and dynamic pruning during inference to reduce memory and improve speed, while the original paper proposes replacing pruned experts with lightweight novices (unbiased estimations). These are distinct compression approaches for MoE models.

Contribution 2: Expert redundancy evaluation using access frequency and output variance

Description: The method introduces a dual-metric approach to evaluate expert redundancy by combining expert access frequency (how often experts are selected) and output variance (stability of expert outputs across calibration data), enabling more accurate identification of redundant experts compared to frequency-based methods alone.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Knowledge partitioning: Context-dependent use of expertise

URL: View paper

Brief Assessment

Knowledge Partitioning[52] studies human expert behavior in bush fire prediction, examining context-dependent knowledge use and performance limitations. It does not address computational expert redundancy metrics in mixture-of-experts models or machine learning systems.

---

2. Joint Spatiotemporal Models for the Estimation of Prey Consumption and Predatorâ□□Prey Overlap: Dynamics of Pacific Cod Predation on Snow and Tanner Crab in the â□¦

URL: View paper

Brief Assessment

Prey Consumption Models[51] focuses on spatiotemporal ecological modeling of predator-prey dynamics in marine ecosystems, not machine learning expert pruning or redundancy evaluation in mixture-of-experts models.

Contribution 3: Robust structured pruning across multiple dimensions

Description: The method demonstrates robust and effective compression performance across three critical dimensions (model architectures, calibration data sources, and calibration sample sizes) where existing structured pruning methods show suboptimal and unstable degradation, achieving up to 2.72 improvement in average zero-shot accuracy.

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. Drpruning: Efficient large language model pruning through distributionally robust optimization

URL: View paper

Brief Assessment

DRPruning[65] focuses on distributionally robust optimization to address uneven performance degradation across domains during pruning, while the original paper addresses robustness across model architectures, calibration data sources, and calibration sample sizes through expert-specific pruning with novice replacement. These are fundamentally different approaches to achieving robustness in structured pruning.

---

2. Structured pruning adapters

URL: View paper

Brief Assessment

Structured Pruning Adapters[59] focuses on parameter-efficient adaptation methods combined with structured pruning for image recognition tasks, not on evaluating robustness across model architectures, calibration data sources, and calibration sample sizes for MoE models as claimed in the original paper.

---

3. One-cycle Structured Pruning with Stability Driven Structure Search

URL: View paper

Brief Assessment

One Cycle Pruning[56] focuses on efficient one-cycle training frameworks for structured pruning with stability-driven search, not on demonstrating robustness across model architectures, calibration data sources, and calibration sample sizes as systematically evaluated in the original paper.

---

4. DepGraph: Towards Any Structural Pruning

URL: View paper

Brief Assessment

DepGraph[58] focuses on generalizing structural pruning across different model architectures (CNNs, RNNs, GNNs, Transformers) through dependency modeling, not on robustness across calibration data sources and sample sizes for MoE models specifically.

---

5. ZipLM: Inference-Aware Structured Pruning of Language Models

URL: View paper

Brief Assessment

ZipLM[63] focuses on inference-aware structured pruning with runtime speedup guarantees for BERT and GPT models, not on robustness across model architectures, calibration data sources, and sample sizes for MoE models specifically. The technical approaches and model families differ fundamentally.

---

6. NIRVANA: Structured pruning reimagined for large language models compression

URL: View paper

Brief Assessment

NIRVANA[57] focuses on NTK-guided pruning with adaptive sparsity allocation and calibration data selection for LLMs, not on demonstrating robustness across model architectures, calibration data sources, and sample sizes as the original paper's core contribution.

---

7. Hydra: Pruning adversarially robust neural networks

URL: View paper

Brief Assessment

Hydra[61] focuses on pruning adversarially robust neural networks using importance score optimization with scaled initialization, targeting adversarial robustness metrics. The original paper addresses structured pruning robustness across model architectures, calibration data sources, and calibration sample sizes for MoE models, which is a fundamentally different problem domain and technical approach.

---

8. Cstar: towards compact and structured deep neural networks with adversarial robustness

URL: View paper

Brief Assessment

Cstar[60] focuses on low-rank tensor decomposition for adversarial robustness, not structured pruning robustness across architectures and calibration data. The candidate addresses a different compression technique (tensor decomposition vs. expert pruning) and different robustness concern (adversarial attacks vs. calibration stability).

---

9. Using Structured Pruning to Find Winning Lottery Tickets

URL: View paper

Brief Assessment

Lottery Tickets Pruning[64] focuses on applying structured pruning to CNNs to find lottery ticket subnetworks, not on evaluating robustness across model architectures, calibration data sources, and sample sizes for MoE models as claimed in the original paper.

---

10. Moreaupruner: Robust pruning of large language models against weight perturbations

URL: View paper

Brief Assessment

MoreauPruner[62] focuses on robustness against weight perturbations (e.g., format changes between bfloat16 and float16) rather than robustness across model architectures, calibration data sources, and sample sizes. The technical approaches differ fundamentally: MoreauPruner uses Moreau envelope-based gradient smoothing, while the original paper uses expert access frequency and output variance metrics for MoE models.

Appendix: Text Similarity Detection

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

References

- [0] MoNE: Replacing Redundant Experts with Lightweight Novices for Structured Pruning of MoE View paper
- [1] Cluster-Driven Expert Pruning for Mixture-of-Experts Large Language Models View paper
- [2] EAC-MoE: Expert-selection aware compressor for mixture-of-experts large language models View paper
- [3] Domain-Specific Pruning of Large Mixture-of-Experts Models with Few-shot Demonstrations View paper
- [4] A survey on inference optimization techniques for mixture of experts models View paper
- [5] Mixture Compressor for Mixture-of-Experts LLMs Gains More View paper
- [6] Not All Experts are Equal: Efficient Expert Pruning and Skipping for Mixture-of-Experts Large Language Models View paper
- [7] Mixture of experts (moe): A big data perspective View paper
- [8] SlimMoE: Structured Compression of Large MoE Models via Expert Slimming and Distillation View paper
- [9] Samoyeds: Accelerating MoE Models with Structured Sparsity Leveraging Sparse Tensor Cores View paper
- [10] ToMoE: Converting Dense Large Language Models to Mixture-of-Experts through Dynamic Structural Pruning View paper
- [11] Efficient Expert Pruning for Sparse Mixture-of-Experts Language Models: Enhancing Performance and Reducing Inference Costs View paper
- [12] Delta Decompression for MoE-based LLMs Compression View paper
- [13] Dense-to-sparse gate for mixture-of-experts View paper
- [14] MC#: Mixture Compressor for Mixture-of-Experts Large Models View paper
- [15] Mixture of Length and Pruning Experts for Knowledge Graphs Reasoning View paper
- [16] STUN: Structured-Then-Unstructured Pruning for Scalable MoE Pruning View paper
- [17] From Sparse to Soft Mixtures of Experts View paper
- [18] Task-Specific Expert Pruning for Sparse Mixture-of-Experts View paper
- [19] Diversifying the Expert Knowledge for Task-Agnostic Pruning in Sparse Mixture-of-Experts View paper
- [20] A Provably Effective Method for Pruning Experts in Fine-tuned Sparse Mixture-of-Experts View paper
- [21] MoE-IÂ²: Compressing Mixture of Experts Models through Inter-Expert Pruning and Intra-Expert Low-Rank Decomposition View paper
- [22] Seer-moe: Sparse expert efficiency through regularization for mixture-of-experts View paper
- [23] Dynamic Mixture of Experts for Adaptive Computation in Character-Level Transformers View paper
- [24] Mosaic Pruning: A Hierarchical Framework for Generalizable Pruning of Mixture-of-Experts Models View paper
- [25] REAP the Experts: Why Pruning Prevails for One-Shot MoE compression View paper
- [26] Edge-moe: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts View paper
- [27] Sparse-MoE-SAM: A Lightweight Framework Integrating MoE and SAM with a Sparse Attention Mechanism for Plant Disease Segmentation in Resource â View paper

- [28] Efficient Expert Pruning for Pre-Training of Mixture-of-Experts Large Language Models View paper
- [29] R^ 2MoE: Redundancy-Removal Mixture of Experts for Lifelong Concept Learning View paper
- [30] Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation View paper
- [31] Scaling Vision with Sparse Mixture of Experts View paper
- [32] Dropping Experts, Recombining Neurons: Retraining-Free Pruning for Sparse Mixture-of-Experts LLMs View paper
- [33] Finding Fantastic Experts in MoEs: A Unified Study for Expert Dropping Strategies and Observations View paper
- [34] Llama-moe v2: Exploring sparsity of llama from perspective of mixture-of-experts with post-training View paper
- [35] Towards Efficient Mixture of Experts: A Holistic Study of Compression Techniques View paper
- [36] Moe-pruner: Pruning mixture-of-experts large language model using the hints from its router View paper
- [37] Demix layers: Disentangling domains for modular language modeling View paper
- [38] Flame: Fully leveraging moe sparsity for transformer on fpga View paper
- [39] Sparse Mixture-of-Experts for Compositional Generalization: Empirical Evidence and Theoretical Foundations of Optimal Sparsity View paper
- [40] Condense, Don't Just Prune: Enhancing Efficiency and Performance in MoE Layer Pruning View paper
- [41] PuzzleMoE: Efficient Compression of Large Mixture-of-Experts Models via Sparse Expert Merging and Bit-packed inference View paper
- [42] Structural adaptation in mixture of experts View paper
- [43] Dense2MoE: Restructuring Diffusion Transformer to MoE for Efficient Text-to-Image Generation View paper
- [44] StructPrune: Structured Global Pruning asymptotics with GPU Memory View paper
- [45] Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate View paper
- [46] ReMod: Learning Structured Sparsity with ReLU Modulation View paper
- [47] Efficient Multimodal Streaming Recommendation via Expandable Side Mixture-of-Experts View paper
- [48] Sparse mixers: Combining moe and mixing to build a more efficient bert View paper
- [49] DiEP: Adaptive Mixture-of-Experts Compression through Differentiable Expert Pruning View paper
- [50] Mixture of Neuron Experts View paper
- [51] Joint Spatiotemporal Models for the Estimation of Prey Consumption and Predatorâ□□Prey Overlap: Dynamics of Pacific Cod Predation on Snow and Tanner Crab in the â□¦ View paper
- [52] Knowledge partitioning: Context-dependent use of expertise View paper
- [53] Unveiling super experts in mixture-of-experts large language models View paper
- [54] A closer look into mixture-of-experts in large language models View paper
- [55] MoE-LLaVA: Mixture of Experts for Large Vision-Language Models View paper
- [56] One-cycle Structured Pruning with Stability Driven Structure Search View paper
- [57] NIRVANA: Structured pruning reimagined for large language models compression View paper
- [58] DepGraph: Towards Any Structural Pruning View paper
- [59] Structured pruning adapters View paper
- [60] Cstar: towards compact and structured deep neural networks with adversarial robustness View paper
- [61] Hydra: Pruning adversarially robust neural networks View paper
- [62] Moreaupruner: Robust pruning of large language models against weight perturbations View paper
- [63] ZipLM: Inference-Aware Structured Pruning of Language Models View paper
- [64] Using Structured Pruning to Find Winning Lottery Tickets View paper
- [65] Drpruning: Efficient large language model pruning through distributionally robust optimization View paper