Adaptive Pruning for Efficient Deep Learning Models

Tech ID: 34826 / UC Case 2026-715-0

BRIEF DESCRIPTION

A novel hardware-software co-designed adaptive pruning technique that improves both efficiency and accuracy of large deep learning models.

FULL DESCRIPTION

This technology introduces a semi-structured pruning method thatretainsthe hardware efficiency of structured pruning while significantly recovering accuracy lost in conventional methods. This co-designed approach combines minor hardware changes with advanced software optimization to generateoptimalpruning masks, enabling large language models (LLMs) and other deep learning architectures tooperatemore efficiently without sacrificing performance. CAP-1D and CAP-2D variants target both inference and training phases, with CAP-2D enabling reusability of pruning masks during training to reduce perplexitysubstantially comparedto current standards.

SUGGESTED USES

- Optimization of large language models (LLMs) for faster and more efficient inference.
- Deployment of resource-efficient AI models in data centers and edge devices.
- Accelerating training pipelines for sparse deep learning networks.
- Integration into modern GPU architectures toleveragehardware-supported pruning.
- Energy-efficient AI solutions for cloud computing and AI-as-a-Service platforms.

ADVANTAGES

- Preserves hardware efficiency benefits ofstructured pruning.
- Recovers up to 99.9% of accuracy lost compared to unstructured pruning.
- Requires negligible hardware modifications involving simple multiplexer wiring changes.
- Enables practical throughput gains supported by modern GPU architectures.
- Supports efficient, deployable, andaccuratesparsity for large-scale models.
- Reduces training perplexity in transposable mask pruning by up to 71%.
- Significantly faster pruning mask generation compared to prior methods.

CONTACT

Alvin Viray
aviray@uci.edu
tel: 949-824-3104.

OTHER INFORMATION

CATEGORIZED AS

- **Computer**
  - Other
  - Software
- **Engineering**
  - Other

RELATED CASES

2026-715-0

UCI Beall
Applied Innovation

5270 California Avenue / Irvine,CA
92697-7700 / Tel: 949.824.2683

© 2026, The Regents of the University of California
Terms of use
Privacy Notice