Paper Title: HOW GRADIENTS SHAPE PRUNING DECISIONS IN LARGE LANGUAGE MODELS

Paper Link: Beyond Size: How Gradients Shape Pruning Decisions in Large Language Models | Papers With Code

1) Summary

1.1 Motivation: To address the increasing computational demands of Large Language Models (LLMs), our motivation stems from the need to develop efficient pruning techniques that maintain performance while reducing parameters and computational costs.

1.2 Contribution: Our work introduces GBLM-Pruner, a novel sparsity-centric pruning method leveraging informative gradients from pre-trained LLMs, outperforming existing counterparts in benchmarks and laying a foundation for advancements in this domain.

1.3 Methodology: GBLM-Pruner operates in a training-free manner by utilizing normalized gradients from calibration samples to determine the importance pruning score, revealing post-pruning structural patterns inherent in LLMs' parameter structure.

1.4 Conclusion: Extensive evaluations demonstrate that GBLM-Pruner surpasses magnitude pruning, Wanda, and SparseGPT, showcasing its effectiveness in identifying sparse networks directly from pre-trained LLMs without the need for retraining or weight updates.

2) Limitations

2.1 First Limitation: One limitation of our approach lies in its reliance on a few calibration samples for gradient-based pruning, which may lead to sensitivity to sample-specific characteristics and affect generalizability.

3) Synthesis: In synthesizing our findings, GBLM-Pruner emerges as a promising, simple, and interpretable solution for LLM pruning, offering an effective balance between maintaining model performance and reducing computational overhead in real-world applications.