Enhancing Large Language Model Efficiency Through Pruning Techniques

Lyes Benacer, Paul Hartmann, Nicolas Schmitt

1. Project description

Context & Motivation
Large language models (LLMs), such as OpenAI's GPT, Facebook's OPT and Meta's Llama, represent significant advancements in AI but require massive computational resources. This results in high energy consumption and considerable carbon emissions, raising environmental concerns. Our project aims to address this issue by exploring ways to reduce the size of LLMs while preserving their performance, making them more efficient and environmentally friendly.

Project Objective
Our primary goal is to prune LLMs (selectively removing certain parameters without retraining them) to maintain model quality while reducing size. By doing this, we aim to improve the sustainability and practicality of deploying these models, especially in resource-constrained environments.

Why It's New & Interesting
While model pruning itself is not entirely new, recent studies have introduced more effective methods for reducing model size with minimal impact on performance. Applying these advanced techniques specifically to large language models (LLMs) represents an emerging and promising area of research, offering unique opportunities for enhancing model efficiency. Pruning addresses the significant deployment challenges LLMs face due to their high resource demands, providing a pathway to make these models more accessible and environmentally friendly. Additionally, by reducing the computational requirements of LLMs, pruning can meaningfully decrease their carbon footprint, contributing to a more sustainable and responsible approach to AI technology in the long term.

Methodology
We start with a base LLM and apply pruning techniques from recent research to assess their impact on performance. Various pruning configurations will be employed to remove non-essential weights and parameters, referencing current studies. To evaluate the pruned models, benchmark tests will compare performance metrics such as accuracy and efficiency against the original model. The goal is to identify an optimal pruned version that maintains accuracy while reducing memory and computational demands, iterating through different pruning levels to find the best size-performance balance.

Real-World Use Case
Consider a real-life deployment scenario where a smaller LLM could operate efficiently on local devices in remote regions with limited internet access and computational resources. By making LLMs smaller, they can be used more widely, particularly in applications like real-time translation, healthcare diagnostics, or educational tools in areas with constrained infrastructure.

2. Project background

In exploring the existing research, it's clear that pruning has been an important technique in machine learning for some time, and more recent work specifically targets large language models (LLMs). Here's how these studies relate to our project:

1. Early Pruning Research by Yann LeCun:
Yann LeCun's 1989 paper, "Optimal Brain Damage" was one of the first to propose the concept of pruning by removing unnecessary weights from neural networks. While this work focused on smaller neural networks, the core idea of reducing model size while preserving performance laid the groundwork for modern pruning techniques. Our project builds on this early concept by applying pruning specifically to large-scale models, aiming to reduce their size and computational cost without a significant drop in performance.

2. Recent Research on Pruning for LLMs:
Recent studies, such as the paper "Pruning Large Language Models: A Survey" (2023) or "https://arxiv.org/pdf/2306.11695" (2024), focus specifically on pruning methods tailored for LLMs. These papers discuss advanced pruning strategies that optimize LLMs by removing less important weights and parameters, improving efficiency without major performance losses. Our project directly connects to this research by applying these cutting-edge techniques to LLMs, testing how pruning can make LLMs more efficient without sacrificing accuracy or capabilities.

3. Benchmarking LLMs:
The article "Benchmarking Large Language Models" from Medium (2023) highlights the importance of rigorous benchmarking when evaluating the performance of pruned models. It provides methods for assessing both the computational efficiency and accuracy of pruned models. This is highly relevant to our project, as we plan to perform similar benchmark tests to measure how well our pruned LLMs perform compared to the original. These benchmarks will help us identify the best pruning strategies that provide the best trade-off between size and performance.

How Our Project Relates to These Studies:
While pruning has been studied in smaller models, the unique challenge with LLMs is their scale and complexity. By applying recent pruning techniques to Facebook's OPT (and potentially other LLMs), our project aims to explore the specific challenges and benefits of pruning large-scale models. Through benchmarking, we will evaluate which pruning methods offer the best balance between reducing size and maintaining performance, thus contributing to the ongoing work on efficient deployment of LLMs in real-world applications. What differentiates our work is the focus on assessing the environmental impact of pruning LLMs, specifically in terms of energy consumption and carbon footprint. We are going to estimate the environmental impact of the pruned models compared to the original ones, an aspect that has not been explored in the referenced studies.

3. Project steps

Project Steps and Methodology

1. **Complete the state-of-the-art review:**
   Research recent pruning techniques for LLMs, focusing on methods and their effectiveness to identify the most relevant approaches.
2. **Apply various pruning techniques:**
   Implement pruning by removing non-essential weights and parameters using techniques like magnitude-based pruning, weight sparsity, or structured pruning.
3. **Study the impact of pruning levels on model performance:**
   Experiment with different pruning levels and measure their effect on performance metrics like accuracy, efficiency, and robustness.
4. **Conduct benchmark tests:**
   Compare the pruned models to the original using performance metrics such as answers's accuracy, processing time, and resource utilization.
5. **Identify the optimal balance between performance loss and cost reduction:**
   Find the pruning configuration that minimizes performance degradation while reducing model size and computational costs.
6. **Estimate the energy consumption and environmental impact:**
   Compare the energy consumption and carbon footprint of the pruned models to the original model to assess environmental impact.
7. **Explore future advancements in pruning and test additional LLMs:**
   Investigate new pruning methods and test them on different LLMs, expanding the research to explore how pruning techniques affect various architectures.

4. First results

Our first experiments, with a 30% pruning ratio, delivered the following:

Inference Time: Reduced slightly from 14.28s to 14.16s, with only a 0.99% improvement.
Model Size: Achieved a 30% reduction in model parameters.

Key Observations:

Our initial 30% pruning resulted in only a minimal speed improvement of 0.99%, indicating that parameter reduction alone does not guarantee computational efficiency. This suggests that magnitude pruning may not be ideal for faster inference. Additionally, only inference time was measured; output quality and ecological impact remain unexamined. With this limited speedup, a 30% pruning ratio may not yield meaningful performance gains.

Next Steps:

We plan to try structured pruning for better hardware acceleration and implement memory tracking to assess system impact. To capture a fuller picture, we'll add quality metrics (e.g., perplexity, loss) and measure energy consumption, alongside expanding benchmarks for varied input types. We'll also investigate alternative and dynamic pruning methods to retain model quality and adapt to real-time needs.

The first results reveal the complexity of balancing model size reduction with practical speed gains, pointing to the need for refined pruning strategies that align efficiency, performance, and sustainability.

Additional content:

Github repo: https://github.com/schmittnicolas/pruning-llm-nlp-project

Références:

1. LeCun, Y. (1989). *Pruning convolutional neural networks*. In *Proceedings of the Neural Information Processing Systems Conference (NeurIPS)*, 1989. Retrieved from https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf

2. Mocanu, D. C., et al. (2023). *Efficient pruning of large language models: A systematic study*. *arXiv:2306.11695*. https://arxiv.org/pdf/2306.11695

3. Zhang, M., et al. (2024). *A survey on the state-of-the-art pruning techniques for large models*. *arXiv:2406.00030v1*. https://arxiv.org/html/2406.00030v1

4. Pei, Y. (2023). *Benchmarking large language models for practical deployment*. Alan. Retrieved from https://medium.com/alan/benchmarking-large-language-models-1e1ab5b809ac