Efficiency and Adaptability through Compression of Large Language Models

Abstract for the general public

In recent years, the rapid growth of large language models (LLMs) has driven major advances in natural language understanding and generation. However, this progress comes with steep computational and memory costs during both training and deployment. Despite their impressive capabilities, state-of-the-art LLMs remain difficult to reproduce, widely inaccessible, and environmentally taxing due to their sheer size. As models continue to scale, the need for effective compression strategies becomes increasingly urgent not only to ease deployment but also to democratize research beyond large industrial labs. A clear example of this challenge is the newly built xAI factory, a powerful computing center designed for training next-generation AI models. This facility alone is estimated to consume hundreds of megawatt-hours of electricity every day - enough to power tens of thousands of homes. The scale of this energy use highlights why it's so important to make AI not only smarter but also more sustainable.

To address this issue, our research will focus on better understanding compression mechanisms and how they affect the scaling laws that govern efficient large language model design. We aim to develop improved compression techniques and conduct experiment-driven studies that will provide insights into a deeper understanding of compression trade-offs in model efficiency and performance. The proposed project contributes to a more accessible and sustainable future for large-scale language models. It will result in methods and insights useful for both academic and industrial communities working on model optimization, interpretability, and sustainable AI development.

Projected Compression

Projected Compression is a novel method that complements standard pruning by taking advantage of all parameters of the pruned model during retraining. Unlike conventional compression approaches that permanently discard pruned weights, Projected Compression retains access to the full set of base model parameters through learnable projection modules. This allows for gradient-based recovery from the original model while preserving training efficiency and compatibility similar to standard transformer architectures. Reintegrating all parameters of the base model during retraining is expected to significantly improve pruning efficiency by allowing more effective reuse of the model's representational capacity.

LLM Logit-Space Compression via Dimensionality Reduction in Distillation

While token-level output probabilities (logits) represent rich semantic information, their full dimensionality (usually more than 30,000 tokens) poses significant memory and communication costs during knowledge distillation. To reduce it, we propose compressing full-vocabulary logits using techniques such as PCA. By distilling in a low-rank logit subspace and expanding only when needed, this approach will reduce hardware overhead while retaining performance. Additionally, by analyzing how individual logit dimensions affect student learning, we gain insight into which tokens carry the most transferable knowledge, improving both the interpretability and distillation process.

Scaling Laws for LLM Hybrid Compression Methods

To guide practical compression choices across use cases and hardware constraints, we will investigate the scaling behavior of hybrid compression methods - those that combine techniques like pruning, quantization, or distillation. Although various compression methods offer different positive trade-offs, their interaction effects are poorly understood. This research will involve both empirical experimentation and theoretical extrapolation modeling to uncover how efficiency scales with model size, quality, sparsity levels, and used compression methods. The goal is to derive generalizable scaling laws and best practices that inform optimal compression strategy choices in the development of large language models.