Exploring Structured Pruning Techniques for Efficient Language Models on Edge Devices

Abhinav Agarwal
abhinav4@stanford.edu
&
Casey Nguyen
caseyhn@stanford.edu

October 8, 2024

Project Category: Natural Language

Project Proposal

Motivation

Large Language Models (LLMs) have achieved remarkable success in various natural language processing tasks. However, their substantial computational and memory requirements pose significant challenges for deployment on resource-constrained devices such as mobile phones and edge devices. With the increasing demand for on-device AI processing—driven by considerations of privacy, latency, and offline functionality—there is a pressing need for efficient language models that maintain high performance while operating within limited hardware capabilities.

Recent advancements like Sheared-LLaMA [1] and LLaMA 3.2 [2] have shown that structured pruning can effectively reduce model size by removing entire structural components (e.g., layers, attention heads) without significant loss in performance. This project aims to explore and develop effective recipes for structured pruning of language models, focusing on adapting pruning techniques to modern architectures like LLaMA 3.1 8B. Our goal is to produce smaller, high-performing LLMs suitable for deployment in edge applications, thereby bridging the gap between powerful NLP models and practical, real-world use cases.

Method

We plan to investigate targeted structured pruning methods tailored to the architectural nuances of LLaMA 3.1 8B, which introduces features such as

Grouped Query Attention (GQA) and larger intermediate (feed-forward network) dimensions (14336 compared to 11008 in LLaMa 2 7B). Our approach involves:

- **Targeted Structured Pruning**:
  - **Pruning Masks**: We will learn pruning masks applied to model parameters at different levels:
    - **Layer-level Masks**: $z_{\text{layer}} \in R^{L_S}$, where L_S is the number of layers in the source model.
    - **Hidden Dimension Masks**: $z_{\text{hidden}} \in R^{d_S}$, where d_S is the hidden dimension.
    - **Head-level Masks**: $z_{\text{head}} \in R^{H_S \times L_S}$, where H_S is the number of attention heads per layer.
    - **Intermediate Dimension Masks**: $z_{\text{int}} \in R^{m_S \times L_S}$, where m_S is the intermediate (FFN) dimension.

    Each mask variable controls whether the corresponding part of the model is pruned or retained. For instance, if $z^i_{\text{layer}} = 0$, the i-th layer is removed.
  - **Constrained Optimization**: We formulate pruning as a constrained optimization problem aiming to minimize the language modeling loss while matching the desired target architecture:

    $$L_{\text{prune}}(\theta, z, \lambda, \phi) = L(\theta, z) + \sum_{j=1}^{L_S} \tilde{L}_{\text{head}_j} + \sum_{j=1}^{L_S} \tilde{L}_{\text{int}_j} + \tilde{L}_{\text{layer}} + \tilde{L}_{\text{hidden}}$$

    where $L(\theta, z)$ is the language modeling loss with pruning masks applied, and $\tilde{L}$ terms are constraint penalties implemented using Lagrange multipliers:

    $$\tilde{L}_{\text{head}_j} = \lambda_{\text{head}} \left(\sum z_{\text{head}_j} - H_T \right) + \phi_{\text{head}} \left(\sum z_{\text{head}_j} - H_T \right)^2$$

    Similar formulations apply to other substructures (layers, hidden dimensions, intermediate dimensions).
  - **Hard Concrete Distributions**: We parameterize the pruning masks using hard concrete distributions to enable learning of discrete prune-or-retain decisions during training.
  - **Handling GQA**: We will adapt our pruning algorithms to account for the GQA mechanism, ensuring that the grouped structure of key/value heads is preserved during pruning. Customized pruning masks will be developed to respect GQA configurations.

  - **Intermediate Dimensions**: Given the larger intermediate dimensions in LLaMA 3.1 8B, we will focus on pruning neurons in the FFN layers, which house a significant portion of the model's parameters.

- **Dynamic Batch Loading**:
  - We will explore dynamic batch loading to adjust data sampling from different domains during continued pre-training, based on loss reduction rates. This aims to optimize data utilization and enhance performance recovery post-pruning.

- **Continued Pre-training and Fine-tuning**:
  - After pruning, we will continue pre-training the model on a language modeling objective using datasets like OpenWebText and WikiText-103 to recover any performance loss.
  - We will fine-tune the pruned model on downstream tasks relevant to edge applications, such as text classification and question answering, using benchmarks like GLUE [3] and SQuAD [4].

Intended Experiments

Our experiments are designed to evaluate the effectiveness of structured pruning on LLaMA 3.1 8B and to understand the trade-offs between model size and performance:

- **Baseline Establishment**:
  - Evaluate the unpruned LLaMA 3.1 8B model on language modeling benchmarks (e.g., perplexity on WikiText-103) and downstream tasks to establish a performance baseline.

- **Pruning Experiments**:
  - **Pruning Ratios**: Experiment with different pruning ratios (e.g., 20%, 40%, 60%) to identify the optimal balance between model size reduction and performance retention.
  - **Component Analysis**: Prune different components (FFN neurons, attention heads, layers) individually and in combination to assess their impact on model performance.
  - **GQA Adaptation**: Test the effectiveness of pruning strategies specifically tailored to handle the GQA structure in attention mechanisms.

- **Performance Evaluation**:
  - **Metrics**: Evaluate using metrics such as perplexity, accuracy, F1-score, and BLEU scores for language tasks.
  - **Efficiency Metrics**: Measure inference latency, memory usage, and throughput on simulated edge environments to assess practical deployment viability.

- **Ablation Studies**:
  - Compare the performance of models pruned using our structured pruning method against those pruned using unstructured pruning and magnitude-based pruning.
  - Analyze the impact of pruning on different layers (e.g., early vs. late layers) and components (e.g., attention vs. FFN layers).
- **Model Analysis**:
  - **Loss Landscape Visualization**: Examine how pruning affects the loss landscape to understand optimization difficulties introduced by pruning.
  - **Attention Patterns**: Analyze attention weights and patterns post-pruning to ensure the model maintains its ability to capture long-range dependencies.
  - **Feature Importance**: Use interpretability techniques (e.g., SHAP values) to understand the contributions of pruned and retained components.
- **Ethical Considerations**:
  - Assess whether pruning introduces or amplifies biases in model outputs by evaluating on datasets designed to test fairness and bias.
  - Discuss implications for user privacy and data security when deploying models on personal devices.

Relevant Dataset and Prior Research

- **Datasets**:
  - **Pre-training Data**: OpenWebText, WikiText-103.
  - **Downstream Tasks**: GLUE benchmark [3], SQuAD [4], and potential edge-relevant tasks like mobile keyboard prediction datasets.
- **Prior Research**:
  - **Sheared-LLaMA** [1]: Demonstrated that targeted structured pruning followed by continued pre-training can produce smaller LLMs with competitive performance using significantly less compute.
  - **LLaMA 3.2** [2]: Showed the feasibility and effectiveness of optimizing LLMs for mobile deployment.
  - **Neural Network Pruning Overview** [5]: Provided comprehensive insights into various pruning techniques and their effectiveness.

Team Members

- **Abhinav Agarwal**: SUNet ID: `abhinav4`
- **Casey Nguyen**: SUNet ID: `caseyhn`

References

[1] Tianle Chen, Zhenheng Tang, Xueguang Ma, Hanxiao Liu, Chunyuan Li, Shujie Liu, Yao Qian, Jianfeng Gao, and Linfeng Song. Sheared Llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023.

[2] Meta AI. LLaMA 3.2: Efficient language models for edge devices. *Meta AI Blog*, 2023.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[5] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of Machine Learning and Systems*, 2020.