OPTIMIZING MULTI-AGENT EFFICIENCY: PARAMETER SHARING AND CLUSTERING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper explores the impact of parameter sharing across layers or agents in a multi-agent framework driven by a large language model. The relevance lies in its potential to enhance training efficiency and model performance, which is crucial as model sizes continue to grow. However, the challenge is identifying which parameters can be shared without degrading performance due to the complexity and diversity of tasks handled by different layers or agents. We solve this by clustering based on activation similarity and employing a shared sub-network approach, enabling effective parameter reuse within clusters. Our method was rigorously tested for training speed, memory usage, and model performance against a baseline without parameter sharing. Results show significant improvements in efficiency with comparable performance, highlighting our approach's balance of trade-offs and its advantage over techniques like pruning and quantization.

1 INTRODUCTION

The emergence of multi-agent frameworks driven by large language models (LLMs) has reshaped various applications, including natural language processing and complex decision-making tasks. Enhancing the efficiency and performance of these frameworks is crucial, especially as model sizes continue to grow. Our research focuses on parameter sharing across different layers or agents, a technique that has shown promise in improving computational efficiency and reducing memory footprint.

Sharing parameters in LLMs and multi-agent systems presents significant challenges. Layers within an LLM, or agents within a multi-agent system, are often specialized for different tasks, making naive parameter sharing potentially detrimental to performance. The core difficulty lies in identifying which parameters can be shared without compromising the unique functionality of each layer or agent.

We address this challenge through a two-pronged approach:

- **Clustering based on activation similarity:** We group layers or agents exhibiting similar activation patterns. This clustering helps identify components that can share parameters effectively.
- **Shared sub-network approach:** Within each cluster, we implement a shared sub-network to reuse parameters across the layers or agents. This method ensures parameter relevance while enhancing overall efficiency.

Our method's effectiveness is validated through extensive experiments focusing on training speed, memory usage, and model performance. We benchmark our parameter-sharing approach against a baseline model without parameter sharing, providing insights into efficiency gains and trade-offs.

OUR CONTRIBUTIONS

- Development of a clustering approach based on activation similarity to identify optimal conditions for parameter sharing.

- Implementation of a shared sub-network approach within clusters to facilitate effective parameter reuse.
- Comprehensive evaluation of our method's impact on training speed, memory usage, and model performance.
- Detailed analysis comparing our approach to traditional efficiency-enhancing techniques like pruning and quantization, highlighting trade-offs and benefits.

Our results show significant improvements in efficiency and performance, positioning parameter sharing as a viable alternative to traditional methods. Future research will explore extending this approach to other architectures and investigating dynamic parameter sharing mechanisms that adapt in real-time during training.

2 Related Work

Parameter sharing has emerged as a promising technique to enhance computational efficiency and reduce memory footprint in large language models (LLMs) and multi-agent systems (MASs). By reusing parameters across layers or agents, these models can achieve scalable and efficient training. For example, **?** explored automation in scientific discovery, emphasizing the potential of shared parameters to reduce training time and enhance performance. Our method builds on these ideas by clustering based on activation similarity to better identify opportunities for parameter sharing.

Several other studies have also explored parameter sharing in neural networks. **?** discussed various techniques and their applications, often assigning unique parameters to each layer or agent. In contrast, our method strategically shares parameters within clusters identified by similar activation patterns, aligning with **?** but adding a more dynamic and granular approach. This ensures that shared parameters remain effective across diverse tasks.

Clustering for efficiency has also been widely used in neural networks (**?**). Studies like **?** employed clustering to address computational challenges, but our method focuses on activation similarities rather than structural attributes. This makes our approach more adaptable to the specific needs of LLMs and MASs, dynamically grouping components based on data-driven patterns.

Finally, traditional memory optimization techniques like pruning and quantization have been well-researched (**?**). While these methods reduce model complexity and computational load, they often compromise accuracy. Our parameter-sharing approach via clustering maintains high accuracy levels while significantly reducing memory usage and training time, positioning our method as a balanced alternative to pruning and quantization.

3 Background

Significant advancements in large language models (LLMs) and multi-agent systems (MASs) are transforming natural language processing and complex decision-making tasks. Pioneering works such as **?** on autonomous scientific discovery and **?** on non-linear dynamics have been instrumental in these advancements, demonstrating the versatility of machine learning algorithms in managing diverse and intricate tasks crucial to our methods and frameworks.

Enhancing the efficiency of LLMs and MASs involves the challenge of sharing parameters across layers and agents without performance loss. Techniques like pruning and quantization have been explored (**?**), but often trade-off between efficiency and accuracy. Our approach uses clustering to maintain performance while optimizing resource usage, focusing on parameter sharing in a clustering framework.

3.1 Problem Setting

We address optimizing multi-agent frameworks in LLMs by strategically sharing parameters. Let $L = \{l_1, l_2, \ldots, l_n\}$ be the layers in an LLM, and $A = \{a_1, a_2, \ldots, a_m\}$ be the agents in a MAS. Each layer or agent has unique parameters P_{l_i} or P_{a_j}. The goal is to find a clustering function C that groups layers or agents with similar activation patterns to share parameters effectively without performance degradation.

An unusual assumption in our model is that high activation similarity among layers or agents allows for effective parameter sharing. This aligns with insights from neuroscientific studies suggesting similar tasks can leverage common neural pathways, contrasting traditional methods that treat each layer or agent as requiring unique parameters.

Our framework also introduces dynamic adaptability in parameter sharing. We propose a mechanism for dynamical parameter allocation based on real-time activation similarities during training, aiming to further enhance model efficiency and adaptability.

4 METHOD

This section details the methodology for implementing parameter sharing across layers or agents in a multi-agent framework driven by a large language model (LLM). Our approach involves clustering based on activation similarity and implementing a shared sub-network to facilitate efficient parameter sharing.

4.1 CLUSTERING BASED ON ACTIVATION SIMILARITY

To identify which layers or agents can share parameters effectively, we analyze their activation patterns. Activation similarity is quantified using the Pearson correlation coefficient between the activation vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ of layers l_i and l_j:

$$S_{ij} = \frac{\text{cov}(\mathbf{a}_i, \mathbf{a}_j)}{\sigma(\mathbf{a}_i)\sigma(\mathbf{a}_j)} \tag{1}$$

Layers or agents with high similarity scores are grouped into clusters. The clustering threshold is a key hyperparameter that ensures a balance between cluster granularity and parameter homogeneity.

4.2 SHARED SUB-NETWORK IMPLEMENTATION

For each cluster C_k containing layers $l_{i_1}, l_{i_2}, \ldots, l_{i_p}$, we establish a shared sub-network $\mathcal{S}_k$ where parameters are shared among these layers:

$$P_{l_{i_1}} = P_{l_{i_2}} = \cdots = P_{l_{i_p}} = P_{\mathcal{S}_k} \tag{2}$$

Shared parameters are optimized collectively for all layers or agents within the cluster, based on their activation similarities. This ensures that parameter sharing does not degrade the performance of specialized layers or agents.

4.3 ADVANTAGES AND IMPLEMENTATION DETAILS

The shared sub-network reduces memory usage and enhances training efficiency by reusing parameters across similar layers or agents. This compact model speeds up training and decreases computational load. We modified the backpropagation algorithm to update shared parameters using the combined gradient contributions from all layers within a cluster.

4.4 EVALUATION METRICS AND BASELINE

We evaluate our parameter-sharing method on three main criteria: training speed, memory usage, and model performance. Comparisons are made against a baseline model with unique parameters for each layer or agent. This highlights the trade-offs and benefits of our approach over the baseline.