COMP4801 Detailed Project Plan

Mixture of Low-Rank Adaptation Pairs of Large Language Models

Liheng Chen
Supervised by Prof. Chuan Wu

1. Project Background

1.1 Large Language Models (LLMs)

Large Language Models (LLMs) have exhibited exceptional capabilities across various natural language processing tasks. Pre-trained on vast datasets, these models can be fine-tuned for specific tasks with remarkable efficiency, achieving state-of-the-art performances. Their deployment spans sectors including healthcare, finance, education, etc. (Solaiman et al., 2023).

LLMs are increasingly being applied in healthcare for tasks such as medical data analysis, clinical decision support, and patient interaction through conversational agents (Solaiman et al., 2023). These systems have the potential to streamline diagnostic processes, personalize patient care, and reduce administrative overhead. Models need to be fine-tuned to handle sensitive medical data with precision and care to avoid misdiagnoses or inequitable treatment suggestions.

In the financial sector, LLMs are deployed for automated customer service, fraud detection, market trend analysis, and financial forecasting (Solaiman et al., 2023). The ability to analyze and process large datasets enables banks and financial institutions to automate decision-making processes and optimize customer interactions. Nonetheless, the complexity of financial data requires careful calibration of these models to avoid issues such as the perpetuation of biases in credit scoring or unfair loan approvals.

Besides playing a transformative role in both the healthcare and financial industry, LLMs also serve as educational tools by providing personalized tutoring systems, content generation for educational material, and even grading systems (Solaiman et al., 2023). These models can help educators deliver customized learning experiences at scale. However, the reliance on AI for grading or feedback could introduce bias, and AI-driven systems might struggle to account for the nuances of student performance in creative or qualitative tasks.

1.2 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT) methods, like P-tuning (Liu et al., 2022), Prefix-tuning (Li & Liang, 2021), and Low-Rank Adaptation (LoRA) (Hu et al., 2021), are designed to minimize the number of trainable parameters needed for model customization. These methods retain most of the model's parameters fixed and only introduce lightweight adaptations, thus significantly reducing the memory and computation required for fine-tuning. LoRA is particularly popular as it decomposes weight matrices into low-rank matrices, enabling efficient updates without fine-tuning the entire model. LoRA reparameterizes the weight matrices of the transformer layers using low-rank decompositions. Several variants of LoRA have emerged to improve its efficiency:

- **VeRA** (Kopiczko et al., 2023) implements inter-layer sharing of two frozen random LoRA matrices while training detached activation vectors for each transformer layer. However, it requires an excessively high rank to be effective because of the limited capacity of models.

- **Tied LoRA** (Renduchintala et al., 2023) is inspired by VeRA to unfreeze shared matrices to be trainable. However, they tie down projection matrices in their settings, restricting their applicability to linear layers of different dimensions.

- **PRoLoRA**: (Wang et al., 2024) showcased enhanced parameter efficiency by sharing sub-matrices within a single linear layer in the transformer, thus avoiding the limitations of both VeRA and Tied LoRA. However, its mechanism concentrates on intra-layer sharing which restricts its potential performance gain in the context of inter-layer sharing.

Overall, these approaches typically assume matrices as the fundamental units for granted and concentrate exclusively on inter-layer or intra-layer sharing, both limiting the potential for parameter efficiency.

2. Project Objective

The primary objective of this project is to develop and evaluate the Mixture of Low-Rank Adaptation Pairs (MoP). This new method combines LoRA with the Mixture of Experts (MoE) (Jacobs et al., 1991) framework. Specifically, MoP aims at:

- Develop and implement a parameter-efficient finetuning method optimized to support a large number of customized models simultaneously while clearly minimizing GPU memory usage.
- Examine overarching sharing principles emphasizing the trade-off between shared components and task-specific differentiation to improve parameter efficiency cost-effectively.
- Achieve parameter efficiency comparable to or better than existing PEFT methods like LoRA, VeRA, Tied LoRA, and ProLoRA.

3. Project Methodology

To achieve the objectives, the project will proceed with the following methodology:

1. **Literature Review**: A detailed review and analysis of current PEFT techniques, focusing on existing limitations in parameter efficiency and task adaptation.

2. **Model Design**: Implement the MoP architecture by integrating LoRA with MoE-like routing mechanisms to select different low-rank adaptation pairs based on the task. The design will involve shard privatization, pair dissociation, and other differentiation strategies to maintain high parameter efficiency while avoiding performance degradation.

3. **Implementation**:

- Baseline Models: Implement baseline models using existing LoRA variants (VeRA, Tied-LoRA, ProLoRA) to serve as benchmarks.
- MoP Development: Refine and tailor the MoP framework for implementation on a specific dataset selection, ensuring seamless integration with established LLM fine-tuning frameworks.

4. **Empirical Evaluation**: Experiments using the LLaMA (Touvron et al., 2023) or GPT (Achiam et al., 2023) series will be conducted to compare the performance of MoP with baseline models. The evaluation will include tasks such as factual knowledge, reasoning, and multilingual tasks. Metrics such as parameter efficiency (memory usage) and task-specific performance (accuracy, reasoning ability) will be recorded.

5. **Ablation Study**: Perform an ablation study to identify the contribution of each component (e.g., pair dissociation, shard privatization) to overall model performance and efficiency.

6. **Experiment Analysis**: Analyze the experimental results to identify areas where MoP outperforms baseline methods, and compile findings into a comprehensive academic essay.

4. Project Schedule and Milestones

Milestone	**Description**	**Completion Date**
Literature Review	Review of existing work on LLMs, PEFT, and LoRA techniques	December 2024
MoP Design	Finalize the MoP architecture and differentiation strategies	December 2024
MoP Implementation	Build the MoP model and integrate it with baseline frameworks	January 2025
Experimentation	Run experiments to evaluate MoP on baseline models	March 2025
Ablation Study	Analyze the impact of individual components on performance	April 2025
Final Analysis and Reporting	Compile results and prepare the presentation	April 2025

5. References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). Gpt-4 technical report. *arXiv Preprint arXiv:2303. 08774*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv Preprint arXiv:2106. 09685*.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(1), 79–87.

Kopiczko, D. J., Blankevoort, T., & Asano, Y. M. (2023). VeRA: Vector-based Random Matrix Adaptation. *arXiv Preprint arXiv:2310. 11454*.

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68.

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *arXiv [cs.CL]*. http://arxiv.org/abs/2101.00190

Renduchintala, A., Konuk, T., & Kuchaiev, O. (2023). Tied-Lora: Enhancing parameter efficiency of LoRA with weight tying. *arXiv Preprint arXiv:2311. 09578*.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., Daumé, H., III, Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., … Subramonian, A. (2023). Evaluating the social impact of generative AI systems in systems and society. In *arXiv [cs.CY]*. http://arxiv.org/abs/2306.05949

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., & Bhosale, S. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv Preprint arXiv:2307. 09288*.

Wang, S., Xue, B., Ye, J., Jiang, J., Chen, L., Kong, L., & Wu, C. (2024). PRoLoRA: Partial Rotation Empowers More Parameter-Efficient LoRA. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2829–2841). Association for Computational Linguistics.