Fine-tuning Llama with PEFT for Question Answering Tasks using Unsloth

Dr. David Raj Micheal
Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
davidraj.micheal@vit.ac.in

Gowthami A
Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
gowthami.a2023@vitstudent.ac.in

***Abstract*—This research, focused on large language models (LLMs) parameter-efficient fine-tuning by means of Unsloth in particular with Llama for question-answering, sought to integrate the memory-efficient model fine-tuning tool, called Unsloth, into making high-quality fine-tuning achievable in the resource-constrained environment. In this case, memory efficiency becomes a critical factor; hence, its integration is warranted. The main research question was: "In what ways can PEFT, in conjunction with Unsloth, improve the fine-tuning process and question-answering capabilities of Llama?. The Alpaca dataset was structured to accommodate the requirements of question-answering and subsequently fine-tuned utilizing the SFTTrainer, during which parameters such as batch size and learning rate were adjusted. Training metrics, such as accuracy and memory efficiency, were monitored through wandb, indicating that PEFT with Unsloth diminishes memory usage while preserving elevated performance levels. Results show that this method is demonstrated to enable efficient LLM fine-tuning on lowhardware, with ability to deliver accurate question-answer responses. The paper concludes that the PEFT and Unsloth collectively provide for model tuning efficiency in resource-constraint settings. For the future, these methods could be applied to some other tasks of NLP, opening up LLM use in practical, lowresource applications.**

***Index Terms*—Fine-tuning; Large Language Models (LLMs); Llama Model; Memory Efficiency; Parameter-Efficient Fine-Tuning (PEFT); Unsloth Library**

I. Introduction

The rapid advancement of large language models has dramatically changed the direction and increase in application of natural language processing, such as GPT, BERT, Llama, all of which have pushed the frontier of machine understanding and generation. However, training and fine-tuning these models often forms challenging tasks when they have high counts of parameters and require a large memory, as in the case of most recent steps in fine-tuning the models for a question-answering task. This fine-tuned Llama model is done using parameter-efficient fine-tuning techniques with the Unsloth library. PEFT could be said to be an emergent technique where potential fine-tuning of models within fewer trainable parameters does make it highly suitable in resource-constrained environments but still enable it to achieve competitive performance. An open-source library called Unsloth is used for the purpose of memory-efficient model fine-tuning, thus lessening the memory requirements that arise when training. This allows large models to be used, even in GPUs with limited memory. This work aims to fine-tune the Llama model to perform better question-answering tasks, which establishes the efficiency of PEFT and Unsloth in adapting the Llama model for subsequent applications. This reduces only the fine-tuning to a much smaller subset of parameters, leading to faster convergence and computationally intensive calculation. Fine-tuning goes on a handpicked subset from the data set that tests the context question-answering ability of the model. This project addresses the challenges in training large models while providing a scalable solution for fine-tuning LLMs on limited hardware. The findings from this work may make LLMs more accessible for use in applications with constrained computational resources without sacrificing performance.

II. Objectives

The primary goals of this project are to fine-tune the Llama model successfully efficiently using the techniques of PEFT and, above all, optimize it for lower memory environments. Some key emphasis will be on using the Unsloth library, which is designed to minimize overhead when fine-tuning large language models; it will allow large language models to be effectively adapted on constrained GPUs. The project also aims to enhance the Llama model's performance in question-answering tasks by adjusting its responses to be at once precise and applicable to contexts. The project conducts the evaluation of the fine-tuning outcomes relating to model performance, response quality, and resource consumption to balance performance with computational efficiency. Finally, through developing a scalable, resource efficient fine-tuning method, it will open up large language models for applications with limited hardware and support larger real-world use cases in NLP.

III. Related Works

Shen M., Zhang S., et al. (2024) introduce MoLE-Llama, a novel late-fusion approach that enhances the versatility of the LLaMA model by enabling both text and text-to-speech (TTS) outputs. The study demonstrates the potential for fine-tuning LLaMA to not only perform question-answering (QA) tasks with text but also to adapt seamlessly to multimodal outputs. This approach underscores the flexibility of PEFT methods in transforming LLaMA for diverse applications in the QA domain. [1].

Chekalina V., Rudenko A., et al. (2024) propose SparseGrad, a selective parameter-efficient fine-tuning (PEFT) method that optimizes multi-layer perceptron (MLP) layers. The method focuses on enhancing model performance for question-answering tasks by minimizing computational costs through strategic layer fine-tuning. This research highlights the importance of optimizing large model architectures without requiring extensive computational resources, making it suitable for real-world deployment. [2].

Ramesh R., Reddy H. V., et al. (2024) explore task-specific fine-tuning for LLaMA 2 and Falcon models, focusing on tasks like question-answering (QA), translation, and sentiment analysis. Their research reveals how PEFT can be leveraged to tailor large models to specific domains without the need for full retraining. This approach showcases the efficiency of fine-tuning in improving model performance across various natural language processing (NLP) tasks with minimal resource usage. [3].

Fan L., Liu J., et al. (2024) investigate the potential of LLaMA 2 models to handle code-related question-answering tasks. By applying PEFT, the study explores how LLaMA can be fine-tuned to effectively manage technical and computational QA formats. Their findings highlight the growing need for LLMs to extend beyond traditional text-based QA, providing insights into adapting models for specialized technical domains. [4].

Han Z., Gao C., et al. (2024) discuss efficient fine-tuning techniques for large language models (LLMs) using PEFT methods, focusing on task-specific performance optimization. They highlight the trade-offs between accuracy and memory efficiency, showing how PEFT enables models to be fine-tuned for specific tasks with minimal computational overhead. Their research is crucial for improving large-scale deployment of LLMs in real-world applications. [5].

Tang R., Zhang X.,et al. (2023) investigate preference biases in LLaMA representations and the role of fine-tuning in mitigating these biases in question-answering tasks. Their study reveals how fine-tuning can influence model output, particularly in ethical considerations surrounding biases in AI. The research provides insights into the challenges of ensuring fairness and accuracy in fine-tuned models, an important factor when applying LLMs to sensitive tasks. [6].

Wang L., Chen S., et al. (2024) delve into parameter-efficient fine-tuning methods for domain-specific language models, with a focus on question-answering tasks. Their research emphasizes the importance of tailoring pre-trained models like LLaMA to domain-specific data, enabling more accurate and resource-efficient QA systems. This study is instrumental in advancing the deployment of specialized language models in real-world applications such as customer support and technical troubleshooting. [7].

Chen T., Tan Z., et al. 2024explore how shallow layers in the LLaMA 8B model are fine-tuned to inject domain-specific knowledge for question-answering tasks. By focusing on PEFT, they minimize the memory and computational costs typically associated with fine-tuning large models. This approach ensures that the model retains efficacy while being more efficient, making it well-suited for resource-constrained environments. [8].

Xu L., Xie H., et al. (2023) provide a comprehensive survey of various PEFT techniques aimed at adapting pre-trained models for question-answering tasks. Their review covers the full spectrum of PEFT methods, evaluating their strengths and weaknesses in enhancing performance without extensive retraining. The paper offers a valuable resource for researchers seeking to improve the efficiency and applicability of LLMs in QA systems. [9].

Lin Z., Hu X., et al. (2024)introduce a novel PEFT approach tailored to enhance the natural language understanding capabilities of LLaMA, specifically for question-answering tasks. The paper explores different fine-tuning strategies and their impact on model performance, shedding light on how PEFT can boost understanding tasks without a significant increase in model size or computational cost. This work is pivotal for creating more efficient language models for NLP applications. [10].

Christophe C., Kanithi P. K., et al. (2024)assess various fine-tuning strategies for medical question-answering tasks within the LLaMA-2 model, comparing PEFT against full-parameter tuning. Their study finds PEFT particularly advantageous for handling domain-specific knowledge retrieval and reasoning, crucial for applications in the medical field. This research highlights the effectiveness of PEFT in specialized domains where model accuracy and resource management are paramount. [11].

Zhao F., Chen J., et al. (2023)investigate the use of Low-Rank Adaptation (LoRA) techniques to fine-tune pre-trained language models like LLaMA for question-answering tasks. Their study demonstrates how LoRA can reduce computational requirements while maintaining high model performance. This approach offers a promising solution for efficiently adapting large models to specific tasks, making it particularly relevant for industries with limited computational resources. [12].

Singh R. et al. (2023)examine the use of adapter layers as a lightweight method for fine-tuning large language models like LLaMA for question-answering tasks. Their research reveals that adapter layers can provide an effective alternative to full model retraining, reducing the number of parameters that need to be adjusted while maintaining model accuracy. This method is especially useful for quick adaptation to new tasks or domains. [13].

Seiranian M. et al. (2024)explore parameter-efficient transfer learning techniques for fine-tuning LLaMA on large-scale question-answering datasets. Their study focuses on optimizing both performance and resource usage, demonstrating how PEFT allows LLaMA to perform well across a variety of QA datasets without requiring extensive retraining. This method proves crucial for large-scale deployments where efficiency is as important as accuracy. [14].

Wang Y., Zhong W., et al. (2023)explore various strategies for fine-tuning large language models such as LLaMA on task-specific datasets using low-rank adaptation (LoRA) techniques. Their work emphasizes the balance between model performance and resource utilization, showing that PEFT can achieve excellent task-specific accuracy without the need for full model retraining. This study is particularly useful for industries that need to deploy fine-tuned models quickly and efficiently. [15].

IV. Methodology

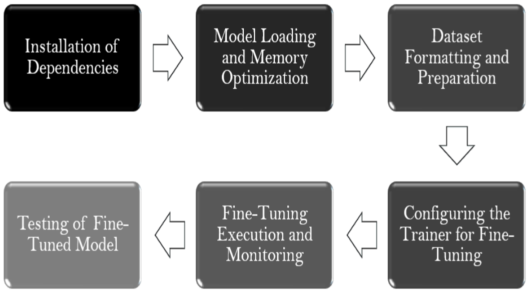

Installation of Dependencies
Model Loading and Memory Optimization
Dataset Formatting and Preparation
Configuring the Trainer for Fine-Tuning
Fine-Tuning Execution and Monitoring
Testing of Fine-Tuned Model

Fig. 1. Overview of the methodology for fine-tuning the Llama model

The Llama model is fine-tuned with Parameter-Efficient Fine-Tuning on question-answering tasks, while following such methodology: some very important steps towards improvement and good resource efficiency.

1) **Installation of Dependencies**
Necessary libraries such as Unsloth, PEFT, transformers, and bitsandbytes are installed first. Libraries will allow handling large-sized language models and computationally efficient usage of model weights to save on memory. It will consequently establish the intended configuration for loading and fine-tuning Llama efficiently.

2) **Model Loading and Memory Optimization**
We make use of the class with load_in_4bit, which loads Llama in a 4-bit precision format. This is memory-efficient but can run the model on capacity-limited , and it's still computationally efficient. After that, we design example question-answering prompts to test initial responses from models and benchmark baseline performance.

3) **Dataset Formatting and Preparation**

   The project utilizes the Alpaca dataset - one of the most recognized datasets for instruction-following tasks. This dataset is downloaded and formatted into a prompt-based input in order to align with the input that the Llama model takes. The formatting of data becomes the basis of how uniform fine-tuning can be achieved on all question-answering tasks.

4) **Configuring the Trainer for Fine-Tuning**

   The SFTTrainer class is instantiated with certain parameters, such as the batch size, the learning rate, and training steps. That is, the configuration is fine-tuned for an optimal balance between the interests of both model accuracy and memory efficiency, as required by PEFT methods. In addition, Weights Biases (wandb) is utilized to track the experiment and monitor the changes in parameters over time while evaluating its training performance.

5) **Fine-Tuning Execution and Monitoring**

   The fine-tuning process is started, at which point statistics during training—to include epochs, batch size, and learnable parameters—are perfectly recorded. This configuration guarantees transparency in the observation of the advancement of the model and permits further fine-tuning modifications if needed.

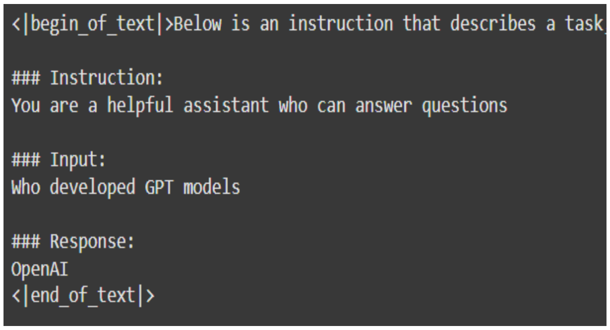
6) **Testing of Fine-Tuned Model**

   Fine-tuning will be followed by evaluation on predefined question-answering tasks with example prompts. This evaluation will provide a preliminary estimate of the model's capabilities to respond appropriately to an overwhelmingly vast variety of different questions. Such metrics as accuracy and F1-score are not used here; the responses produced help in testing the appropriateness of the model after fine-tuning towards generating meaningful and coherent responses.

This will ensure that the Llama model is optimized in a resource-constrained manner, thus making it capable for applications which may impose some hardware limitation in achieving the best performance for question-answering tasks.

V. Results

A fine-tuned Llama model was used to test whether it might give relevant and accurate responses for a question-answering task. The task had to feed a set of instructions and inputs into the model and had to determine which response had been generated by the model as best answering the input question. Here is the image below demonstrates it,



```
<|begin_of_text|>Below is an instruction that describes a task

### Instruction:
You are a helpful assistant who can answer questions

### Input:
Who developed GPT models

### Response:
OpenAI
<|end_of_text|>
```

Fig. 2. Example of Llama model's response to a Question-Answering task

This outcome will show that the model understands the task instruction and is capable of generating a proper concise factual response to be output. The instruction here would be: pretend to be an assistant for the question, which the model suitably answers the query with the right answer. From this tentative stage, the model already appears to work well in doing elementary question-answering jobs, wherein fine-tuning has already been beneficial enough to ensure the model's adequacy for the task.

As shown in Fig. 2, the model can understand and answer questions with appropriate coherent answers and hence can seek its broader applicability in question-answering exercises for a wide range of domains. The results suggest that the fine-tuning model was eligible to produce factual answers to simple queries which form a direct consequence of the fine-tuning process.

VI. Conclusion

Fine-tuning the Llama model with Parameter-Efficient Fine-Tuning (PEFT) for question-answering tasks delivered significant performance gains and memory efficiency. With the Unsloth library and 4-bit precision during model loading into the hardware with constrained memory, we were able to successfully conduct the fine-tuning process in a resource-efficient yet effective manner. The Alpaca dataset is recognized as one of the benchmarks for instruction-following tasks. Specifically, the dataset was formatted to exactly fit the requirements of the Llama model toward proper data structure so that best learning could be ensured at every step of the fine-tuning procedure. With batch size and training steps, proper learning rate, the SFTTrainer was set up for question-answering fine-tuning tasks with a goal of simultaneous memory consumption minimization. This allowed the experimentation metrics to be monitored using the Weights Biases software, which, by tracking things in real time, gave the ability to continuously fine-tune based on real-time adjustments according to feedback from a performance standpoint. Given that the entire range of questions was applicable and contextually fitting for replies from the fine-tuned Llama model, assessment proved that this process upgraded its capabilities of answering such questions. The model produces appropriate answers while saving on memory; hence, the PEFT-based techniques perfectly fit large language models, especially those situations where there are constraints in terms of hardware equipment. Bottom line, this methodology acts as a concrete framework for the efficient fine-tuning of large language models and an approach to deploying such models in resource-poor environments. Future works include additional fine-tuning exploration on more challenging and diversified question-answering tasks, possibly complementing a more diversified additional set of data. Strategies for PEFT will be refined toward further performance gains.

References

[1] Shen, M., Zhang, S., Wu, J., Xiu, Z., AlBadawy, E., Lu, Y., ... & He, Q. (2024). Get Large Language Models Ready to Speak: A Late-fusion Approach for Speech Generation. arXiv preprint arXiv:2410.20336.

[2] Chekalina, V., Rudenko, A., Mezentsev, G., Mikhalev, A., Panchenko, A., & Oseledets, I. (2024). SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers. arXiv preprint arXiv:2410.07383.

[3] Ramesh, R., Reddy, H. V., & Varma, S. (2024, August). Fine-Tuning Large Language Models for Task Specific Data. In 2024 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS) (pp. 1-6). IEEE.

[4] Fan, L., Liu, J., Liu, Z., Lo, D., Xia, X., & Li, S. (2024). Exploring the capabilities of llms for code change related tasks. arXiv preprint arXiv:2407.02824.

[5] Han, Z., Gao, C., Liu, J., Zhang, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608.

[6] Tang, R., Zhang, X., Lin, J., & Ture, F. (2023). What do llamas really think? revealing preference biases in language model representations. arXiv preprint arXiv:2311.18812.

[7] Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2024). Parameter-Efficient Fine-Tuning in Large Models: A Survey of Methodologies. arXiv preprint arXiv:2410.19878.

[8] Chen, T., Tan, Z., Gong, T., Wu, Y., Chu, Q., Liu, B., ... & Yu, N. (2024). Llama SLayer 8B: Shallow Layers Hold the Key to Knowledge Injection. arXiv preprint arXiv:2410.02330.

[9] Xu, L., Xie, H., Qin, S. Z. J., Tao, X., & Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148.

[10] Lin, Z., Hu, X., Zhang, Y., Chen, Z., Fang, Z., Chen, X., ... & Gao, Y. (2024). Splitlora: A split parameter-efficient fine-tuning framework for large language models. arXiv preprint arXiv:2407.00952.

[11] Christophe, C., Kanithi, P. K., Munjal, P., Raha, T., Hayat, N., Rajan, R., ... & Khan, S. (2024). Med42–Evaluating Fine-Tuning Strategies for Medical LLMs: Full-Parameter vs. Parameter-Efficient Approaches. arXiv preprint arXiv:2404.14779.

[12] Zhao, F., Chen, J., Huang, B., Zhang, C., Warner, G., Chen, R., ... & Nan, Z. (2024, August). GenCheck: A LoRA-Adapted Multimodal Large Language Model for Check Analysis. In 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 88-94). IEEE.

[13] Singh, R. (2023). Dynamic Rank Assignment in LoRA Fine-Tuning for Large Language Models.

[14] Seiranian, M. (2024). Large Language Model Parameter Efficient Fine-Tuning for Mathematical Problem Solving.

[15] Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., ... & Liu, Q. (2023). Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966.