Optimizing Small Language Models with Task-Specific LoRA Adapters for Superior Performance in Memory-Constrained Environments

Anweasha Saha, Bharati Jagdish Panigrahi, Lance Garrick Soares,
Shaurya Bhatnagar, Vijay Ravichander

September 27, 2024

Abstract

General-purpose LLMs are pre-trained on vast and diverse datasets, designed to handle a wide range of tasks, however, they might lack the specificity needed for highly specialized domains (Paul, September 29, 2023). On the other hand, hosting multiple specialized LLMs on the same hardware requires large computational resources making it infeasible. This document is a project proposal for a Small Language Model composed of multiple trained Low-Rank Adaptation (LoRA) adapters (VLLM, 2024) to boost the model inference capabilities on Mathematical Reasoning, Code Generation, Text Summarization, and Text Paraphrasing tasks by leveraging LoRAX (hot swapping the LoRA adapters) based on the task.

1 Hypothesis

Fine-tuning multiple LoRA adapters with Small Language Model, each specialized for a particular task, will lead to better or comparable performance than general-purpose Large Language Models (LLMs) in memory-constrained environments such as mobile devices and laptops.

2 Related Work

Recent advancements in optimizing language models for resource-constrained environments have shown promising results. One significant approach (Hu et al., 2021) involves the use of low-rank adapters for fine-tuning pre-trained models on specific tasks. This technique substantially reduces the number of trainable parameters, thereby decreasing memory and computational requirements while maintaining performance.

Building upon this concept, the study (Wang et al., 2022) introduced AdaMix, a framework that utilizes a mixture of adapters for parameter-efficient fine-tuning of large language models. AdaMix enhances task adaptability and performance by combining multiple task-specific adapters, all while keeping the number of trainable parameters low.

The practical application of these principles is evident in Apple's recent work on foundation language models (Gunter et al., 2024). Their research introduces two models: AFM-on-device, a 3 billion parameter model designed for efficient on-device use, and AFM-server, a larger cloud-based model. These models power Apple Intelligence features across various platforms, employing task-specific LoRA adapters to tailor performance for different applications. This approach demonstrates the effectiveness of the technique in maintaining efficiency and accuracy in real-world, memory-constrained environments.

3 Methodology

3.1 Base Model

Small Language Model (<3B parameters) - Llama 3.2 (1B/3B) *or* Qwen 2.5 Instruct Models (under 3B) *or* Gemma 2 2B Instruct Model

3.1.1 Comparison Model

Mistral-7B (Jiang et al., 2023)

3.1.2 Datasets

1. **Mathematical Reasoning:**
   allenai/math_qa
   openai/gsm8k

2. **Code Generation:**
   lucasmccabe-lmi/CodeAlpaca-20k
   flytech/python-codes-25k

3. **Summarization and Paraphrasing::**
   ccdv/arxiv-summarization
   ccdv/cnn_dailymail

3.1.3 Milestones:

1. Collect baseline inference metrics for the Mistral 7B model using the evaluation datasets.
2. Collect baseline inference metrics for the base Small Language Model (SLM) using the evaluation datasets.
3. Fine-tune a separate LoRA adapter for each specific target task using the base SLM.
4. Evaluate the base SLM's inference performance on the evaluation datasets by applying the corresponding LoRA adapter for each target task.

3.1.4 Eval Benchmarks:

HumanEval - Code Eval
Subset of Hellaswag - General Eval

4 Inputs and Outputs:

Text Summarization	
Input	**Output**
LONDON, England (Reuters) – Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe(...)	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday. Young actor says he has no plans to fritter his cash away. Radcliffe's earnings from first five Potter films have been held in trust fund.

Text Paraphrasing	
Input	**Output**
Sunsets paint the sky with vibrant hues, marking the transition from day to night. The colors—deep oranges, pinks, and purples—create a moment of calm and beauty, reminding us of nature's rhythm. Each sunset, though fleeting, offers a quiet reflection on the passage of time.	Sunsets fill the sky with vivid colors, signaling the shift from day to night. Shades of orange, pink, and purple bring a sense of peace and beauty, highlighting nature's rhythm. Though brief, each sunset provides a gentle reminder of time's passing.

Code Generation	
Input	**Output**

Create a function that takes a specific input and produces a specific output using any mathematical operators. Write corresponding code in Python.	`def f(x):`
	`"""`
	`Takes a specific input`
	`and produces a`
	`specific output using`
	`any mathematical`
	`operators`
	`"""`
	`    return x**2 + 3*x`

Mathematical Reasoning	
Input	**Output**
Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	Natalia sold 48/2 = «48/2=24»24 clips in May. Natalia sold 48+24 = «48+24=72»72 clips altogether in April and May.

References

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Peebles, Ke Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans, Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao, Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, Anupama Mann Anupama, Areeba Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Dominik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fangping Shi, Felix Bai, Frank Chu, Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey Bigham, Jeffery Cao, Jeff Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega, Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi, Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar, Xin Wang, Xin Zheng, Walker Cheng, Yael Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yunsong Meng, Zhao Tang Luo, Zhi Ouyang, Alp Aygar, Alvin Wan, Andrew Walkingshaw, Andy Narayanan, Antonie Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin, Irina Belousova, Joris Pelemans, Karen Yang, Keivan Alizadeh Vahid, Liangliang Cao, Mahyar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, Regan Poston, Sam Xu, Shuangning Liu, Sushma Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar, Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia, Zhile Ren, and Zhongzheng Ren. 2024. Apple intelligence foundation language models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Anwesha Paul. September 29, 2023. General purpose vs. customizable llms: Weighing in on the debate.

VLLM. 2024. Dynamically serving lora adapters.

Yaqing Wang, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models. *arXiv preprint arXiv:2205.12410*, 1(2):4.