M2: Lora Finetuning

Xueqing Xu (xx823)

Department of Physics, University of Cambridge

April 7, 2025

Introduction

- Brief overview of the coursework objectives
- Introduction to LoRA (Low-Rank Adaptation) and its application to LLMs
- Introduction to time series forecasting using language models
- Mention of the predator-prey system (Lotka-Volterra) as the target application

This coursework explores the application of Low-Rank Adaptation (LoRA) to fine-tune the Qwen2.5-Instruct Large Language Model (LLM) for forecasting predator-prey population dynamics. Building on the observation that LLMs can function as time series forecasters without explicit training[1], we investigate how targeted fine-tuning can enhance their forecasting capabilities while maintaining efficiency.

LoRA represents a parameter-efficient fine-tuning technique that dramatically reduces the number of trainable parameters by injecting small, trainable low-rank matrices into existing model weights without modifying the original parameters. This approach is particularly valuable when working with large models under computational constraints.

The target application is forecasting the Lotka-Volterra predator-prey system, a classic ecological model that describes the dynamic interaction between two species. This system exhibits oscillatory behavior that presents a challenging forecasting task requiring understanding of non-linear dynamics and the ability to model complex interdependencies.

Methodology

Qwen2.5-Instruct model architecture

The Qwen2.5-0.5B-Instruct model implements a decoder-only transformer architecture with 494 million parameters (Table 1). With a hidden dimension of 896 across 24 transformer layers (Table **??**), the model balances depth and computational efficiency.

A key architectural feature is Grouped Query Attention (GQA), which employs 14 query heads but only 2 key-value heads—a 7:1 ratio that substantially reduces memory usage during inference while maintaining attention capabilities. For normalization, Qwen2.5 uses RMSNorm with $\epsilon = 10^{-6}$, which offers better training stability than traditional LayerNorm.

Property	**Value**
Model Name	Qwen2.5-0.5B-Instruct
Model Type	qwen2
Total Parameters	494,032,768 (∼0.5B)
Architecture	Decoder-only Transformer
Precision	bfloat16

Table 1: Qwen2.5-0.5B-Instruct Model Overview

References

[1] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635, 2023.