SCSE23-1095: Evaluating the Carbon Footprint of Code Implementation

Tar Sreeja
College of Computing and Data Science
Nanyang Technological University

Supervised by:
Asst. Prof. Lim Wei Yang Bryan

INTRODUCTION

As large language models (LLMs) gain popularity in natural language research, it becomes imperative to address the computational demand and ecological impact of their fine-tuning processes. This study explores the emissions and energy consumption associated with fine-tuning three LLMs, namely LLaMA-2, Mistral and Gemma, across three GPUs (T4, L4, and A100), for three natural language processing (NLP) tasks — question answering (QA), summarisation, and sentiment analysis. By examining the tradeoffs between performance, emissions and runtime across model configurations, hyperparameters and hardwares, this project aims to provide insights into minimising the carbon footprint of LLM training.

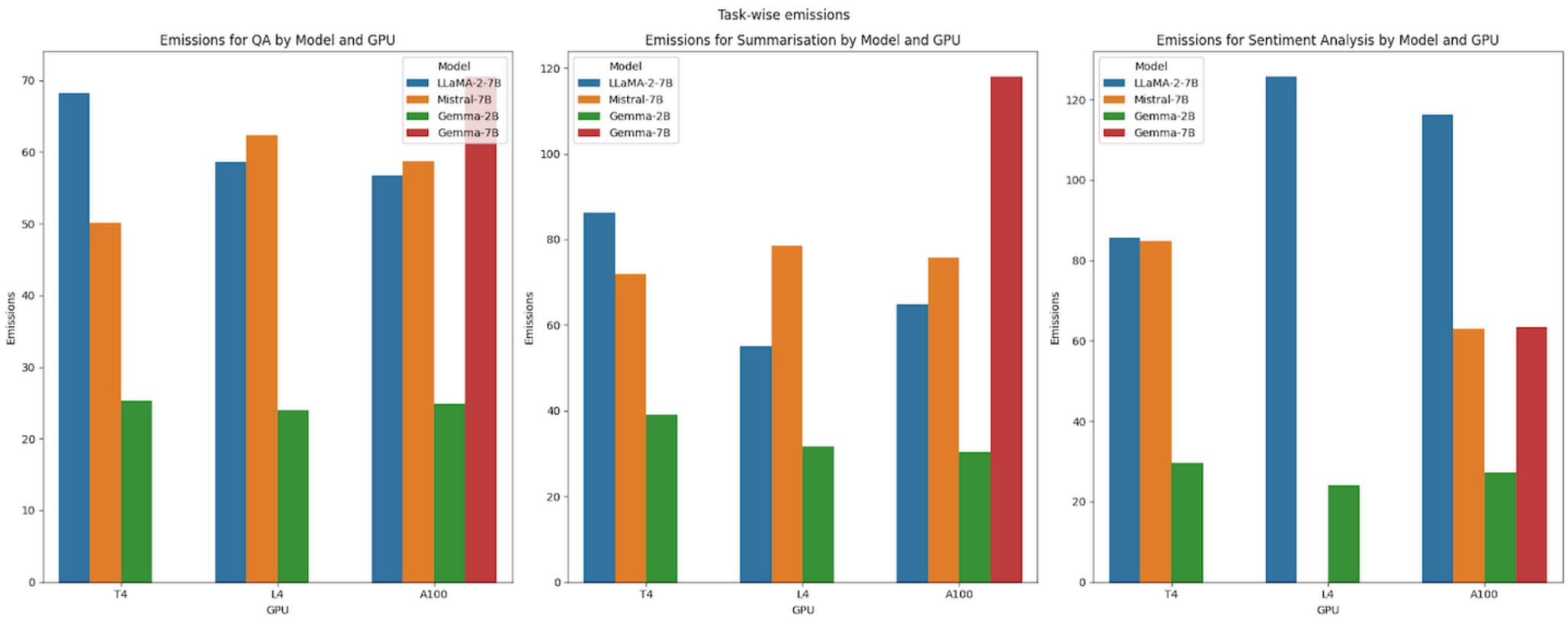
OBJECTIVES

1. Empirical evaluation of the carbon footprint associated with fine-tuning large language models across different tasks and model configurations
2. Providing insights into the variability of emissions across different GPU types, highlighting the importance of hardware selection in environmentally conscious model training
3. Examining the potential of model optimisation to mitigate carbon emissions, proposing practical recommendations for reducing the environmental impact of NLP research

METHODS

Evaluation Package: CodeCarbon
Models: LLaMA-2-7B, Mistral-7B, Gemma-2B, Gemma-7B
Datasets & Tasks: Benchmark datasets across three tasks: QA, summarisation, and sentiment analysis.
Fine-tuning approach: Parameter Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA)
GPUs: T4, L4, and A100, selected for their varied performance and energy efficiency profiles.
Hyperparameter Optimisation: Emissions and validation loss were optimised through multi-objective tuning with Optuna across 10 trials.
Metrics: Emissions (kg CO_2 equivalent), emissions rate (kg CO_2/hour), and total runtime (hours) were recorded.

RESULTS



ANALYSIS AND DISCUSSION

CROSS-GPU COMPARISON

- Emissions Rates: T4 < L4 < A100.
- Total emissions do not follow a similar trend as emissions rates. Less powerful GPUs like T4 require longer runtimes, leading to more energy consumption over time. Conversely, A100 finishes tasks much faster, despite its higher emissions rate per second, resulting in comparable total emissions. In fact, ***total CO2 emitted by A100 is less as compared to the other two GPUs in most cases***.

HYPERPARAMETER OPTIMISATION

- **Batch size**: While smaller batch sizes, such as 2 or 4, were commonly found in the best trials reported by Optuna, further comparative analysis suggests that a batch size of 8 strikes the best balance between performance and emissions efficiency. By selecting slightly larger batch sizes, training iterations become more efficient, reducing the total number of steps required for convergence. This leads to a reduction in the overall resource consumption, which directly correlates to lower carbon emissions.
- **Accumulation steps**: Trials with accumulation steps between 2 and 6 were frequently observed among the best results, with lower accumulation steps being more prominent in ideal configurations.
- Learning rate and warmup steps: No correlation with performance and emissions
- **Epochs**: Trials with fewer epochs, typically 1 or 2, were associated with lower emissions while maintaining competitive validation losses.