ALTA Tutorial: Welcome Letter

Nicholas I-Hsien Kuo
Centre for Big Data Research for Health (CBDRH)
University of New South Wales
`n.kuo@unsw.edu.au`

Dear Participants,

Welcome to the ALTA 2024 Tutorial! This session is designed to explore efficient techniques for training small-scale large language models (LLMs) in resource-constrained environments. As AI capabilities expand, deploying powerful models effectively remains a key challenge. This tutorial will provide practical insights to help overcome these limitations.

Tutorial Overview

The tutorial is divided into six parts, each addressing a key topic:

1. **Part 1: Introducing LoRA with a Simple Example** — Demonstrates Low-Rank Adaptation (LoRA) using a "Delete 4" setup on MNIST to illustrate parameter-efficient adaptation.
2. **Part 2: Quantisation Fundamentals** — Covers mixed-precision arithmetic in PyTorch, highlighting trade-offs between computational efficiency and accuracy.
3. **Part 3: Quantisation Techniques for LLMs** — Explores NF4, GPTQ, and GGUF methods for deploying LLMs on constrained hardware, with practical demonstrations.
4. **Part 4: Advanced Quantisation and Deployment Strategies** — Focuses on INT4 representations and visualisation of quantisation effects to optimise memory usage.
5. **Part 5: Parameter-Efficient Fine-Tuning (PEFT)** — Details techniques like LoRA and 4-bit quantisation applied to models such as LLaMA-2.
6. **Part 6: Implementation and Best Practices** — Integrates prior techniques with best practices for fine-tuning and deployment using Hugging Face's ecosystem.

Tutorial materials can be accessed at: `https://figshare.com/articles/book/Hands-On_NLP_with_Hugging_Face_ALTA_2024_Tutorial_on_Efficient_Fine-Tuning_and_Quantisation/27929580?file=50876241`

Learning Outcomes

By the end of this tutorial, you will:

- Understand core principles of LoRA and quantisation.
- Gain hands-on experience with memory-efficient fine-tuning.
- Be equipped to deploy LLMs on resource-constrained hardware.

We look forward to your participation in unlocking the potential of resource-efficient LLMs!

Best regards,

Nicholas I-Hsien Kuo
Centre for Big Data Research in Health (CBDRH)
The University of New South Wales, Sydney, Australia
`n.kuo@unsw.edu.au`

Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association, pages 213–213
December 2-4, 2024 ©2024 Association for Computational Linguistics