Ziqi Dai

desky_morebetter@bupt.edu.cn | Beijing, China

Education

Beijing University of Posts and Telecommunications (BUPT) Beijing, China
M.S. in Communication Engineering, GPA: 3.92/4.00 Sep 2022 – Jun 2025
Selected Coursework: Machine Learning; Graph Theory

Beijing University of Posts and Telecommunications (BUPT) Beijing, China
B.S. in Communication Engineering, GPA: 3.22/4.00 Sep 2018 – Jun 2022
Selected Coursework: Linear Algebra; Probability Theory; Data Structures & Algorithms; Python Programming

Language & Skills

Language: English (IELTS 6.5)
ML/LLM: SFT; LoRA/QLoRA; Quantization (GPTQ/AWQ, 8/4-bit); KV-cache optimization; FlashAttention; Prompting; Evaluation (F1, PPL, Latency)
Programming & Tools: Python; PyTorch; Hugging Face (Transformers, PEFT, TRL); scikit-learn; pandas; Git; Linux

Research Experience

Parameter-Efficient Fine-Tuning of Large Language Models for Sentiment & Topic Classification
Jan 2025 – Mar 2025
Independent Research (LLM Training) Technologies: LLaMA-2/3, LoRA/QLoRA, Hugging Face (Transformers, PEFT, TRL), PyTorch

- Addressed single-GPU constraints by applying LoRA/QLoRA with 4-bit quantization for sentiment/topic tasks.
- Explored ranks, scaling α, learning-rate schedules, and micro-batch sizes; tracked throughput and validation F1.
- Achieved accuracy/F1 comparable to full SFT with ∼90% reduction in trainable parameters; reproducible logs and seeds.

Efficient Inference of LLMs via Low-bit Quantization and KV-Cache Optimization May 2025 – Aug 2025
Independent Research (LLM Inference Optimization) Technologies: GPTQ/AWQ (8/4-bit), KV-cache tuning, FlashAttention, Hugging Face Transformers

- Built benchmarking harness for throughput/latency profiling (tokens/s, p50/p95 latency) on long-context inference.
- Compared FP16 vs. INT8/INT4 quantization; evaluated perplexity degradation.
- Tested KV-cache compression/eviction and FlashAttention kernels; documented CUDA profiling and reproducible scripts.
- Demonstrated significant speedup and memory reduction with minimal perplexity degradation; summarized deployment recommendations for single-GPU inference.

Fine-tuning BERT for Sentiment Analysis Apr 2024 – May 2024
Technologies: BERT, Python, PyTorch

- Fine-tuned BERT on Twitter/IMDb sentiment datasets; tuned hyperparameters to improve generalization.
- Applied data augmentation in low-resource settings; compared accuracy/F1 with baselines.

LSTM-Based Communication User Equipment State Prediction Feb 2024 – Apr 2024
Technologies: LSTM, 5G RedCap, Python

- Modeled RSRP time series in 5G RedCap devices for predictive handover in smart-factory environments.
- Tuned LSTM depth/hidden size; improved handover success rate while reducing energy consumption.

Pose Recognition and Activity Classification Sep 2023 – Dec 2023
Technologies: YOLOv5-Pose, LSTM/GRU, Python

- Built two-stage pipeline for human-pose recognition and activity classification; curated/cleaned datasets.
- Fine-tuned YOLOv5-Pose and sequence models with cross-validation; improved stability and end-task performance.

Publications

Dai, Z. (2024). *Research on Measurement Relaxation and Predictive Handover Based on LSTM Networks.* In *Proceedings of the 2024 International Conference on Communication Technology (ICCT 2024).* (First Author)

Honors & Awards

Third-class Scholarship (Sep 2020), Progress Award (Sep 2020), First-class Scholarship (Dec 2022), First-class Scholarship (Dec 2023) — BUPT

Thesis

Thesis Title: *Research on Robust Handover Algorithms for 5G RedCap in Power IoT Applications*