Wentao Guo

wentaoguo.me | wg0420@princeton.edu

EDUCATION

Princeton University

- Ph.D. in Computer Science 09/2024 -

Cornell University

- Master of Engineering in Computer Science, GPA: 3.993 06/2022 - 12/2023
- B.S. in Computer Science with Honors, Magna Cum Laude, GPA: 3.890 09/2018 - 05/2022

PUBLICATION & MANUSCRIPT

*** denotes equal contribution.**

- **Wentao Guo**, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, Beidi Chen, Zhaozhuo Xu. "Zeroth-Order Fine-Tuning of LLMs with Extreme Sparsity." *In the WANT workshop at ICML'24.* [paper]
- A. Feder Cooper*, **Wentao Guo***, Khiem Pham*, Tiancheng Yuan, Charlie F. Ruan, Yucheng Lu, Christopher De Sa. "Coordinating Distributed Example Orders for Provably Accelerated Training." *In NeurIPS'23.* (Also in the DMLR workshop at ICML'23) [paper] [poster] [code]
- Yucheng Lu, **Wentao Guo**, and Christopher De Sa. "GraB: Finding Provably Better Data Permutations than Random Reshuffling." *In NeurIPS'22.* [paper] [poster]
- **Wentao Guo***, Andrew Wang*, Bradon Thymes, Thorsten Joachims. "Ranking with Slot Constraints." *In KDD'24.* [paper] [slides] [poster] [code]
- Tao Yu*, **Wentao Guo***, Jianan Canal Li*, Tiancheng Yuan*, Christopher De Sa. "MCTensor: A High-Precision Deep Learning Library with Multi-Component Floating-Point." *In the HAET workshop at ICML'22.* [paper] [poster] [code] [video]
- Yann Hicke, Abhishek Masand, **Wentao Guo**, Tushaar Gangavarapu. "Assessing the efficacy of large language models in generating accurate teacher responses." *In the BEA workshop at ACL'23.* [paper]

RESEARCH EXPERIENCE

Research Assistant

Prof. Beidi Chen's Lab, Carnegie Mellon University 06/2023 - 05/2024

- Memory-efficient LLM fine-tuning on devices [**WANT@ICML'24**]
  - Combined sparse fine-tuning techniques with zeroth-order (ZO) optimization methods to personalize LLM fine-tuning while respecting memory constraints on devices (8 GiB).
  - Demonstrated our sparse fine-tuning method's better performance than ZO full fine-tuning, other sparse ZO fine-tuning baselines, and ZO-PEFT methods as ZO-LoRA and ZO with Prefix Tuning.

Research Assistant

Prof. Christopher De Sa's Lab, Cornell University 06/2021 - 05/2023

- Centralized example ordering for improved optimizer convergence [**NeurIPS'22**]
  - Collaborated to develop the Gradient Balancing (GraB) algorithm that leverages per-example gradients from the prior epoch to determine the example order in the next epoch, with a provably faster convergence rate than the random reshuffling (RR) method.
  - Demonstrated a 40% wall-clock time convergence speedup of GraB over RR and a 68% memory reduction over the data ordering algorithm from prior research in the LeNet classification task.

- Distributed example ordering for improved optimizer convergence [**NeurIPS'23**]
  - Designed the Coordinated Distributed GraB (CD-GraB) algorithm that generalizes the GraB algorithm to the distributed setting without centralized access to all data examples.
  - Collaborated to prove that CD-GraB enjoys a linear speedup in the number of distributed workers, and achieves a faster convergence rate than the distributed RR method.
  - Demonstrated that CD-GraB reduces the training steps by 15% to reach the same test loss in small-scale GPT-2 pretraining task.
- High-precision floating-point computation for learning in hyperbolic space [**HAET@ICML'22**]
  - Developed the MCTensor library that implements high-precision Multiple-Component Format (MCF) algorithms with PyTorch-compatible interfaces, and the HTorch library that integrates hyperbolic space optimization pipelines with MCF algorithms.
  - Showed that MCF models could reduce the error of Poincaré Halfspace embedding tasks by 7%.

Research Assistant
Prof. Thorsten Joachims's Lab, Cornell University 06/2022 - 02/2023

- Ranking with slot constraints [**KDD'24**]
  - Proposed the MatchRank algorithm that recommends a shortlist of relevant candidates while respecting the set of slot constraints defined by decision-makers.
  - Collaborated to prove that MatchRank yields tight approximation guarantees on its ranking objectives.
  - Validated MatchRank's performance on the Cornell admission dataset and analyzed the robustness of MatchRank under the scenario of inaccurate estimation of candidates' relevance level.

TEACHING EXPERIENCE

- **Graduate Teaching Assistant, Cornell**
  - CS 4787 Principles of Large-Scale Machine Learning Systems 08/2023 - 12/2023
  - CS 4780 Intro to Machine Learning 01/2023 - 05/2023

ACADEMIC SERVICE

- **Reviewer**: NeurIPS'23, ICLR'24, ICML'24, KDD'24, NeurIPS'24, Journal of DMLR, ICLR'25, ACL RR (June 2024, October 2024)

HONORS

- Princeton University Graduate Fellowship
- Cornell Engineering Honor Society membership
- Cornell Dean's List for 6 semesters