

Wentao Guo

🏠 wentaoguo.me | ✉ wg247@cornell.edu

RESEARCH INTEREST

Scalable and efficient ML algorithms

EDUCATION

Princeton University

- Ph.D. in Computer Science Aug 2024 -

Cornell University

- Master of Engineering in Computer Science, GPA: 3.993 Jun 2022 - Dec 2023
- B.S. in Computer Science with Honors, Magna Cum Laude, GPA: 3.890 Sep 2018 - May 2022

PUBLICATION & MANUSCRIPT

* denotes equal contribution.

- 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 submission to KDD’24*. [\[paper\]](#) [\[code\]](#)
- 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 Jun 2023 - May 2024

- Memory-efficient LLM finetuning on devices
 - Combined sparse finetuning techniques with zeroth-order (ZO) optimization methods to personalize LLM finetuning while respecting memory constraints on devices (8-12 GiB GPU memory).
 - Demonstrated our sparse finetuning method’s better performance than magnitude-based and random sparse mask, and PEFT methods like LoRA and prefix tuning.

Research Assistant

Prof. Christopher De Sa’s Lab, Cornell University Jun 2021 - May 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% CUDA 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 a 15% training step convergence speedup for CD-GraB in GPT-2 training tasks.
- High-precision floating-point computation for learning in hyperbolic space [**ICML'22 HAET workshop**]
 - 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

Jun 2022 - Feb 2023

- Ranking with slot constraints [**arXiv, in submission to 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.

DEVELOPER EXPERIENCE

• Developer Lead

Pathways Project, Prof. René Kizilcec's Lab, Cornell University

Jun 2021 - May 2023

- Proposed and implemented search algorithms that provide diverse suggestions on course enrollment choices while fitting with students' situational interests, and deployed the **Pathways website**.
- The **Pathways website** serves more than 3000 students in Cornell, with highlights from the Registrar.

• Backend Developer & Tester Lead

Course Management System X (CMSX) Project, Cornell University

Sep 2019 - May 2022

- Fixed tens of production system bugs, contributed more than 11,000 lines of code, reviewed 76 peer's pull requests, and supervised the tester team and 2 external project team's progress.
- The **CMSX website** serves more than 8000 students and faculties over 100 courses in Cornell.

TEACHING EXPERIENCE

• Graduate Teaching Assistant

– CS 4787 Principles of Large-Scale Machine Learning Systems

Aug 2023 - Dec 2023

– CS 4780 Intro to Machine Learning

Jan 2023 - May 2023

• Undergraduate Course Consultant

– CS 3110 Data Structures & Functional Programming

Aug 2021 - Dec 2021

ACADEMIC SERVICE

- **Reviewer:** NeurIPS'23, ICLR'24, ICML'24, KDD'24, Journal of DMLR

HONORS

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