Wentao Guo

wentaoguo.me | wg247@cornell.edu

RESEARCH INTEREST

Efficient ML Algorithms, Distributed Learning, Machine Learning Systems

EDUCATION

Cornell University

- Master of Engineering in Computer Science, GPA: 4.031 — Jun 2022 - Dec 2023
- B.S. in Computer Science with Honors, Magna Cum Laude, GPA: 3.890 — Sep 2018 - May 2022

Tsinghua University

- Non-degree, Cornell Study-away Program, GPA: 3.80 — Sep 2020 - Jan 2021

PUBLICATION & MANUSCRIPT

*** denotes equal contribution.**

- **Wentao Guo***, A. F. Cooper*, Khiem Pham*, Tiancheng Yuan, Charlie F. Ruan, Yucheng Lu, Christopher De Sa. "CD-GraB: 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 SIGIR'24.* [paper] [code]
- **Wentao Guo***, Tao Yu*, 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 - Present

- Zeroth-order optimization with prompt tuning
  - Combining zeroth-order optimization with prompt tuning to further enable memory-efficient LLM finetuning, investigated prompt decomposition heuristics, and performed scaling experiments across language modeling and GLUE tasks.
- Token separation behavior in attention training
  - Investigating the mechanism behind softmax attention's token separation behavior and established its connection to top-k sparse attention with OPT training from scratch tasks.

Research Assistant

Prof. Christopher De Sa's Lab, Cornell University — Jun 2021 - May 2023

- CD-GraB: distributed data ordering **[NeurIPS'23, ICML'23 DMLR workshop]**
  - Designed the CD-GraB algorithm that utilizes a dedicated global order server to coordinate the next-epoch distributed data orderings that would not require data movement across distributed workers.
  - Identified that CD-GraB would preserve the performance when the number of distributed workers increases (linear speedup w.r.t. the number of workers), and CD-GraB would save 15% steps compared to distributed random reshuffling to reach the same perplexity on training GPT-2 from scratch tasks.

- GraB: centralized data ordering **[NeurIPS'22]**
  - Collaborated to develop the GraB algorithm that reorders the data examples for the next epoch via reordering current epoch data examples to minimize the average gradient error.
  - Showcased a 40% wall-clock time convergence speedup of GraB over random reshuffling and a 0.96 GiB memory reduction over the prior offline greedy algorithm in the LeNet classification task.
- MCTensor: high-precision arithmetic for hyperbolic learning **[ICML'22 HAET workshop]**
  - Constructed bottom-to-top Multi-Component Float (MCF) operators for efficient high-precision computation, fused NN modules and optimizers with MCF tensors, and released the MCTensor library.
  - Developed hyperbolic manifold operations with MCF algorithms, and released the HTorch library to enable the numerically-robust hyperbolic space optimization pipeline.

Research Assistant
Prof. Thorsten Joachims's Lab, Cornell University Jun 2022 - Feb 2023

- MatchRank: ranking with slot constraints **[arXiv]**
  - Proposed the MatchRank algorithm that maximizes the Monte-Carlo estimate of the maximum bipartite matching (MBM) size on sampled candidate-slot relevance graphs per ranking step.
  - Proved MatchRank's approximation guarantees, and improved the efficiency of MatchRank by caching the augmenting paths and using faster approximate greedy maximization algorithms.
  - Illustrated the superiority of MatchRank over Probability Ranking Principle heuristics on Cornell admission dataset, and analyzed the robustness of MatchRank on inaccurate relevance estimation.

DEVELOPER EXPERIENCE

- **Developer Lead**
  Pathways Project, Prof. René Kizilcec's Lab, Cornell University Jun 2021 - May 2023
  - Designed search algorithms that provided suggestions on course enrollment choices, developed the backend with Flask, MongoDB, and Redis, and launched the website to serve 3000 Cornell students.
- **Backend Developer & Tester Lead**
  Course Management System, Cornell University Sep 2019 - May 2022
  - Fixed 10s MySQL and Java production bugs on backend, created 75 and reviewed 76 peer's pull requests, and supervised new members, and held weekly meetings to manage the team.
  - The website serves more than 8000 students and faculties over 100 courses in Cornell.
- **Game Development Intern**
  QQ Speed Mobile Team, Tencent, Shenzhen, China Jun 2020 - Aug 2020
  - Programmed game modules in Unity with C#, created tools to accelerate project loading and compilation time and analyzed the performance of C# libraries on serialization and deserialization.

TEACHING EXPERIENCE

- **CS 4787 Principles of Large-Scale Machine Learning Systems**
  Graduate Teaching Assistant, Cornell University Aug 2023 - Dec 2023
- **CS 4780 Intro to Machine Learning**
  Graduate Teaching Assistant, Cornell University Jan 2023 - May 2023
- **CS 3110 Data Structures & Functional Programming**
  Undergraduate Course Consultant, Cornell University Aug 2021 - Dec 2022

ACADEMIC SERVICE

NeurIPS'23, ICLR'24 Reviewer

HONORS

Cornell Engineering Honor Society (Tau Beta Pi), Dean's List for 6 semesters, Honorable Mention in MCM 2018