

# Wentao Guo

🏠 wentaoguo.me | ✉️ wg247@cornell.edu

## RESEARCH INTEREST

---

Model-Efficient ML, Data-Efficient ML, 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, † denotes alphabetical order.)

- A. F. Cooper\*†, **Wentao Guo**\*†, 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 DMLR workshop at ICML’23) [\[paper\]](#) [\[poster\]](#) [\[code\]](#)
- **Wentao Guo**\*, Andrew Wang\*, Bradon Thymes, Thorsten Joachims. “**Ranking with Slot Constraints.**” In *submission to SIGIR’24*. [\[paper\]](#) [\[code\]](#)
- Yucheng Lu, **Wentao Guo**, and Christopher De Sa. “**GraB: Finding Provably Better Data Permutations than Random Reshuffling.**” In *NeurIPS’22*. [\[paper\]](#) [\[poster\]](#)
- 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 *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 *BEA workshop at ACL’23*. [\[paper\]](#)

## RESEARCH EXPERIENCE

---

### Carnegie Mellon University

Jun 2023 - Present

*Research Assistant, Prof. Beidi Chen’s Lab, Carnegie Mellon University*

- **Zeroth-order optimization with prompt tuning**
  - Combined zeroth-order optimization with prompt tuning to save both optimizer states and forward activations for memory, performed scaling experiments across language modeling and GLUE tasks.
- **Token separation behavior in attention training**
  - Investigated the mechanism behind softmax attention’s token separation behavior, established its connection to top- $k$  sparse attention with OPT pretraining tasks.

### Cornell University

Jun 2021 - May 2023

*Research Assistant, Prof. Christopher De Sa’s Lab, Cornell University*

- **CD-GraB: distributed data ordering [NeurIPS’23, DMLR workshop]**
  - Designed the CD-GraB algorithm that utilizes a dedicated global order server to determine the next-epoch distributed data orderings that would not require data movement across distributed workers.
  - Demonstrated that CD-GraB would preserve the performance when the number of distributed workers increases, and CD-GraB would save 15% training steps on pretraining GPT-2 on WikiText-103.

- **GraB: centralized data ordering** [NeurIPS'22]

- Collaborated to develop the GraB algorithm that reorders the data examples for next epoch via reordering current epoch data examples to minimize the average gradient error.
- Showcased the data efficiency of GraB over random reshuffling by 40% wall-clock time convergence speedup in LeNet classification, LSTM language modeling tasks.

- **MCTensor: efficient high-precision arithmetic for hyperbolic learning** [HAET workshop]

- Constructed bottom-to-top Multi-Component Float (MCF) operators for efficient high-precision computation, and fused NN modules and optimizers with MCF tensors. Achieved a 7% error reduction in Poincaré Halfspace embedding tasks by replacing 64-bit PyTorch tensors with MCF tensors.
- Developed hyperbolic manifold operations with MCF tensors, and built hyperbolic NN modules and Riemannian optimizers on top to perform numerically-robust hyperbolic space optimization.

## Cornell University

Jun 2022 - Feb 2023

*Research Assistant, Prof. Thorsten Joachims's Lab, Cornell University*

- **MatchRank: ranking with slot constraints** [arXiv]

- Proposed the MatchRank algorithm that maximizes the Monte-Carlo estimate of the size of maximum bipartite matching (MBM) on sampled candidate-slot relevance graphs per ranking step.
- Improved the efficiency of MatchRank by caching the augmenting paths, parallelizing the Monte-Carlo estimation, and using faster approximate greedy maximization (Stochastic Greedy) algorithm.
- 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** Fall 2023
- **CS 4780 Intro to Machine Learning** Spring 2023
- **CS 3110 Data Structures & Functional Programming** Fall 2021

## ACADEMIC SERVICE

---

- **NeurIPS'23, ICLR'24 Reviewer**

## HONOR

---

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