Wei-Lin Chiang

Email: weichiang@berkeley.edu Webpage: `infwinston.github.io`

Education

Ph.D. in EECS, University of California, Berkeley *Aug. 2020 - present*

- AI systems at Sky Computing Lab, advisor: Prof. Ion Stoica

M.S. in Computer Science Dept., National Taiwan University *Feb. 2018 - Jul. 2020*

- Thesis: efficient algorithms for training deep and large graph convolutional networks
- Advisor: Prof. Chih-Jen Lin, GPA: 4.26/4.3

B.S. in Computer Science Dept., National Taiwan University *Sep. 2013 - Jan. 2018*

- Minor in Mathematics. GPA: 4.06/4.3 (major GPA: 4.17/4.3)

Research Interests

AI systems, sky computing

- SkyPilot: an intercloud broker system for AI on any cloud
  `https://github.com/skypilot-org/skypilot`
- FastChat: an open platform for chat LLMs, powering Vicuna and Chatbot Arena
  `https://github.com/lm-sys/FastChat`

Projects

SkyPilot: an intercloud broker system for AI workloads on any cloud

- Paper published at NSDI'23 and open-source system at GitHub (3.3k stars)
- Adopted by 10+ AI labs and organizations

Vicuna: an open-source chatbot impressing GPT-4 with 90*% ChatGPT quality

- Our Vicuna demo has served over **3 million requests** (Blog, Demo, Weights)
- Total **3 million downloads** from Vicuna models, 600+ vicuna-based models on HuggingFace

FastChat: an open platform for training, serving, and evaluating LLM chatbots

- The project has gained **25k** GitHub stars, Arena demo, LLM leaderboard
- Chatbot Arena has collected **50K** human votes on anonymous comparisons of 20+ chat LLMs (Dataset)

LLM as a Judge: LLM judges for chatbot evaluation with multi-turn chat benchmark (MT-Bench)

- Scalable, effective, and validated benchmark for chat LLMs
- LLM leaderboard, GitHub, Paper in submission to NeurIPS'23

Balsa: a learned query optimizer without expert demonstrations

- Optimize SQL queries by deep RL and sim-to-real learning, matching expert-designed optimizers
- Paper published at SIGMOD'22; Github

Cluster-GCN: an efficient algorithm for training large and deep GCN

- Paper published at KDD'19 with **800+** citations; Github

Work Experience

Intern@Amazon Product Graph, Seattle *May 2021 - Aug 2021*

- Proposed contrastive pre-training techniques for semi-structured data
- Few-shot learning with BERT on information extraction benchmark (SWDE)

- Mentors: Colin Lockard

Intern@Google Research, Mountain View *Dec 2018 - Mar 2019*

- Developed efficient algorithms for training large (million-scale) and deep GCN models
- Achieved state-of-the-art performance on several public datasets (PPI, reddit)
- Mentors: Prof. Cho-Jui Hsieh and Si Si

Intern@Alibaba Group, Hangzhou *July 2017 - Sept 2017*

- Developed distributed ML algorithms on Alibaba's parameter server (KunPeng)
- Reduced the training time (5% ~ 30%) of billion-scale models behind Ads and recommendation systems
- Mentors: Prof. Chih-Jen Lin and Wei Chu

Research Intern@Microsoft, Redmond *July 2016 - Oct 2016*

- Developing large-scale ML algorithms on Microsoft's distributed platform (REEF)
- Implemented Newton's method for solving billion-scale Ads CTR problems
- Mentors: Prof. Chih-Jen Lin and Sathiya Keerthi

Publications & Preprints

- L. Zheng*, **W.-L. Chiang***, S. Ying*, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z Li, D. Li, E. Xing, H. Zhang, J. Gonzalez, I. Stoica, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena" in submission to **NeurIPS'23**
- **W.-L. Chiang**, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. Gonzalez, I. Stoica, E. Xing (alphabetical order), "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality", blog post
- Z. Wu, **W.-L. Chiang**, Z. Yang, E. Friedman, S. Shenker and I. Stoica., "Optimizing Spot Instance Savings under Deadlines,", in submission to **NSDI 2024**
- Z. Yang, Z. Wu, M. Luo, **W.-L. Chiang**, R. Bhardwaj, W. Kwon, S. Zhuang, F. Luan, G. Mittal, S. Shenker and I. Stoica. "SkyPilot: An Intercloud Broker for Sky Computing," **NSDI 2023**
- Z. Yang and **W.-L. Chiang*** and S. Luan* and G. Mittal and M. Luo and I. Stoica. "Balsa: Learning a Query Optimizer Without Expert Demonstrations," **SIGMOD 2022**
- Y.-S. Li*, **W.-L. Chiang***, and C.-p. Lee. "Manifold Identification for Ultimately Communication Efficient Distributed Optimization," **ICML 2020**
- **W.-L. Chiang**, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh. "Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks," **KDD 2019**
- C.-Y. Hsia, **W.-L. Chiang**, and C.-J. Lin. "Preconditioned Conjugate Gradient Methods in Truncated Newton Frameworks for Large-scale Linear Classification," **ACML 2018** (**Best Paper Award**)
- **W.-L. Chiang**, Y.-S. Li, C.-p. Lee, and C.-J. Lin. "Limited-memory Common-directions Method for Distributed L1-regularized Linear Classification," **SIAM SDM 2018**
- **W.-L. Chiang**, M.-C. Lee, and C.-J. Lin. "Parallel Dual Coordinate Descent Method for Large-scale Linear Classification in Multi-core Environments," **KDD 2016**
- M.-C. Lee, **W.-L. Chiang**, and C.-J. Lin. "Fast Matrix-vector Multiplications for Large-scale Logistic Regression on Shared-memory Systems," **ICDM 2015**

Awards and Honors

- **Best Paper Award, ACML** *2018*
- **Bachelor Thesis Award, First Prize, National Taiwan University** *2017*
- **Innovative Undergraduate Research Award, Ministry of Science and Technology** *2017*
- **Undergraduate Research Award, First Prize, NTU CSIE** *2016*