Zifan Liu

https://www.linkedin.com/in/zifan-liu-cs

Email : zifan@cs.wisc.edu
Mobile : (608) 733-9266

Education

University of Wisconsin-Madison, Madison, WI *Sep. 2018 – Feb. 2024 (expected)*
Ph.D. Candidate in Computer Science; GPA: 4.0/4.0 *Advisor: Theodoros Rekatsinas*

Research focus: data-centric AI, data cleaning, data selection, ML model debugging

Shanghai Jiao Tong University, Shanghai, China *Sep. 2014 – Jun. 2018*
B.S. in Computer Science; GPA: 3.92/4.3

Experience

Microsoft, Redmond, WA *May. 2022 – Aug. 2022*
Research Intern at Gray Systems Lab *Host: Shaleen Deep*

- Introduced and implemented an efficient algorithm to verify denial constraints on tabular data, achieving up to **40x speedup** over the state of the art on datasets containing tens of millions of rows.
- Proposed and implemented an anytime algorithm for denial constraint discovery that starts outputting constraints within the first 10 minutes of execution while previous systems need a blocking building phase that takes more than 48 hours.

Google, Mountain View, CA *May. 2021 – Aug. 2021*
Research Intern at TFX *Hosts: Evan Rosen, Paul Suganthan*

- Designed and implemented a novel strategy, utilizing Apache Beam for distributed computation, to identify problematic data slices for ML model debugging. Achieved up to **80% cost reduction**.
- Implemented a feature enabling users to define time constraints for problematic data slice identification.

Projects

Data selection for language model finetuning *Apr. 2023 – Present*

- Designed and implemented an algorithm to curate data from a vast corpus for language model finetuning, given a small set of examples from the target domain.
- Boosted the F1 scores of the domain-specific classification tasks by up to **4 points** by finetuning the model on 1 million examples curated from a corpus of 150 million text sequences.

Picket: a framework to safeguard against data corruptions in ML pipelines *Jan. 2020 – Jan. 2021*

- Developed a framework that identifies and mitigates data corruptions during training and deployment of machine learning models over tabular data, enhancing ML robustness.
- Enhanced the test accuracy by up to **6 points** in downstream ML tasks using the framework when 20% of the training data are poisoned.

Selected Publications

Zifan Liu, Evan Rosen, Paul Suganthan G. C.. AutoSlicer: Scalable Automated Data Slicing for ML Model Analysis. *NeurIPS Workshop on Challenges in Deploying and Monitoring ML Systems, 2022.*

Zifan Liu, Jongho Park, Theodoros Rekatsinas and Christos Tzamos. On Robust Mean Estimation under Coordinate-level Corruption. In *Proceedings of ICML*, 2021.

Zifan Liu, Zhechun Zhou and Theodoros Rekatsinas. Picket: Guarding Against Corrupted Data in Tabular Data during Learning and Inference. *The VLDB Journal*, 2021.

Skills

Programming Languages (Proficient) Python, Java, SQL; (Familiar) C/C++

Tools PyTorch, TensorFlow, Pandas, Numpy, Jupyter Notebook, Docker, Apache Spark & Beam, Git