Shuo Zhang

sz3177@columbia.edu (857) 350-2596 inverse0.github.io shuo-zhang-leo Inverse0

Education

Columbia University *Sept 2023 – Dec 2024*
MS in Computer Science; GPA: 3.95/4.00

Boston University *Sept 2020 – May 2023*
BA in Computer Science and Economics; GPA: 3.83/4.00

- Honors: Magna Cum Laude

Publications

Data-Centric Text-to-SQL with Large Language Models
Zezhou Huang, ***Shuo Zhang***, Kechen Liu, Eugene Wu
3rd Table Representation Learning Workshop @ NeurIPS, 2024

Data Cleaning Using Large Language Models
Shuo Zhang, Zezhou Huang, Eugene Wu
In Sumbission

Policy Enforcement for IoT: Complexities and Emerging Solutions
Shuo Zhang, Luoyao Hao, Henning Schulzrinne
IEEE International Conference on Performance, Computing and Communications (***IPCCC***), 2024

Advancing IoT System Dependability: A Deep Dive into Management and Operation Plane Separation
Luoyao Hao, ***Shuo Zhang***, Henning Schulzrinne
In Sumbission

CAVE: Concurrency-Aware Graph Processing on SSDs
Tarikul Islam Papon, Taishan Chen, ***Shuo Zhang***, Manos Athanassoulis
Proceedings of the ACM on Management of Data (***SIGMOD***), 2024

Research Experience

Research Assistant *New York, NY*
Columbia University Database Group. Advisor: Prof. Eugene Wu & Prof. Kenneth Ross *May 2024 – present*

- Developed an advanced data-cleaning system using Large Language Models (LLMs) to automate and enhance the cleaning process, achieving up to 27% higher F1-scores compared to systems like Holoclean and Baran on benchmark datasets.
- Identified critical flaws in text-to-SQL benchmarks, uncovering up to 37% incorrect ground truth queries, and emphasized the impact of data cleaning in improving conversion accuracy.
- Designed a data-centric text-to-SQL framework that incorporates data preprocessing, cleaning, relationship graph building, and business logic integration, outperforming ground truth by up to 33.89% on BIRD benchmarks.
- Optimized join-aggregate query performance by developing a hybrid approach that dynamically selects between hash joins and sparse/dense matrix multiplication, achieving up to a 6x speedup over DuckDB.

Research Assistant *New York, NY*
Columbia University Internet Real-Time Lab. Advisor: Prof. Henning Schulzrinne *Sept 2023 – present*

- Researched IoT systems and applications, leading to the development of a novel IoT system that enhances dependability through the separation of management and operational planes.
- Introduced the Identity-Independent Policy (IIP), a flexible and expressive policy specification framework, and evaluated its expressiveness using real-world IoT IFTTT datasets, demonstrating up to 99% expressibility.
- Developed a Python-based policy server prototype integrating IIP to demonstrate the system's effectiveness.
- Addressed key challenges in IoT policy enforcement, such as interoperability, policy expressiveness, and ease of creation,

by proposing solutions that leverage WoT, Apple Pkl, and LLMs to simplify and resolve these issues.

Research Assistant *Boston, MA*
Boston University DiSC Lab. Advisor: Prof. Manos Athanassoulis *June 2022 – May 2024*

- Contributed to the design and development of an innovative graph processing system that leverages SSD parallelism.
- Implemented the system in C++ with a suite of parallelized graph traversal algorithms (BFS, DFS, WCC, PageRank, and Random Walk).
- Developed a block variation of CAVE, leveraging caching to avoid block revisiting issues and achieve near-optimal performance on small cache sizes, resulting in an 11x speedup over the non-blocked variant.
- Demonstrated the effectiveness of the system through rigorous testing on diverse real-world graph datasets across multiple SSD platforms, achieving up to 984x speedup compared to existing systems like GraphChi, Mosaic, and GridGraph.

Industry Experience

AI Intern *San Jose, CA*
VisionX LLC *June 2022 – Aug 2022*

- Worked on an AI product designed to identify crop pests and diseases, enhancing detection accuracy and efficiency.
- Implemented an object detection model using PyTorch and the YOLO framework, achieving 98% accuracy and a 180% improvement in processing speed compared to the previous product.
- Fine-tuned the model to address real-world challenges, enhancing performance on blurry and low-resolution images.
- Developed a Java-based Graphical User Interface (GUI) to improve user experience and streamline interaction with the detection system.

Projects

Video Super-Resolution *2024*

- Developed a method that enhances both video resolution and smoothness by integrating advanced image super-resolution techniques with video interpolation strategies, achieving a peak signal-to-noise ratio (PSNR) of 34.72 on a subset of the Vimeo-90K benchmark.

E-commerce Website *2024*

- Developed a shopping website, utilizing HTML for design and Flask for creating a RESTful API that interfaces with a MongoDB backend.
- Implemented SQL to develop essential e-commerce functionalities such as shopping carts and order tracking systems, mirroring those found on established shopping platforms.

Time Series Analysis *2023*

- Researched the potential of machine learning models for time series data forecasting.
- Experimented with various state-of-the-art machine learning models, including Long Short-Term Memory (LSTM) networks, and compared their performance with traditional methods such as the autoregressive integrated moving average (ARIMA).
- Evaluated the models using real-world GDP data, achieving an 85% improvement in Root Mean Squared Error (RMSE) compared to the ARIMA predictions.

Data Cleaning and Integration *2022*

- Collaborated with Boston City Councilor Ruthzee to analyze small landlord data, identifying properties with the potential to join affordable housing programs.
- Applied machine learning techniques, including random forest, to classify owner-occupied and small landlord properties for integration into affordable housing databases.
- Cleaned and consolidated existing housing databases to identify eligible properties, and developed machine learning models to cluster housing units and predict fair rental price ranges.

Technologies

Programming Languages: Python, Java, SQL, C++, HTML/JavaScript, R, OCaml

Technologies: PostgreSQL, MongoDB, MySQL, DuckDB, Flask, Node.js, MQTT, PyTorch, Stata