Zhihao Lyu

(669) 499-7073 | lv.zh@northeastern.edu | Personal Website

EDUCATION

Northeastern University **Sep. 2020 ~ Dec. 2023**

Master of Science in Computer Science, CGPA: 3.9/4.0

- Professional Development Award recipient: funding for nationwide participation in academic conferences
- Research Capstone: Using a Language Model's Perplexity for Evaluating a Trajectory's Outlierness [paper]
- Research Talk : A Semantic Parsing Method for SQL using Language Models with Data Augmentation [slides]

Beijing University of Civil Engineering & Architecture **Sep. 2015 ~ Jun. 2020**

Bachelor of Engineering in Urban Planning, GPA: 3.5/4.0

- Two-time Merit Scholarships recipient: awarded for ranking within the top 10% of academic performance and receiving Excellence Awards in the National Social Survey Competition.

RESEARCH EXPERIENCE

Northeastern University **Aug. 2023 ~ Present**

Research Capstone [paper] | Advised by Mario Nascimento and Michał Aibin

- Leveraged **perplexity** of **large language models** like **BERT** to identify trajectory outliers without relying on labeled data, contributing to the detection of taxi scam or changes in self-driving car routes.
- Constructed a customized vocabulary and corpus using over 50,000 GPS coordinates and created a adaptive mapping, which link coordinates to words using their **spatial distribution and density**.
- Introducing artificially missing and drifting coordinates to augment the training data. Applied 16 × 16 **road network** and **relative position bias** to enhance the model's **spatial awareness**.
- Training with **Transformer** encoder with BERT-style tasks achieved an **~0.8** AUC and ~**0.7** F1 score.
- Calculated perplexity score of unseen trajectory, employing **K-means** for binary classification on scores to identify outliers. F1 score of the prediction results **surpasses the state-of-the-art model** by **9%** to **16%**.
- Utilizing non-augment data and different clustering methods for ablation analysis confirms the superiority.

MIT Transit Lab **Aug. 2023 ~ Present**

Research Assistant | Advised by Haris Koutsopoulos

- Collaborated with **Chicago Transit Authority,** funded by the **Department of Energy**, to address public transportation service reliability issues, employing machine learning algorithms for bus scheduling.
- Established a **cloud computing platform** using Flask, modified distributed transaction for database, and integrated Google Cloud Platform to store and continuously update bus location, speed, and arrival time.
- Utilized Data Frame Algebra in Pandas to merge these **real-time data** and calculate service reliability metrics such as number of passengers, waiting time, and cycle time. Updated bus scheduling strategy every minute, and **offered a web application** for experts to evaluate scheduling strategy using **React.js**.
- Successfully operated the system for three researchers and dozens of dispatchers for months, collecting over 10,000 high-quality data points to improve algorithm performance.

Northeastern University **Jul. 2023 ~ Oct. 2023**

Research Talk [slides] | Advised by Jeongkyu Lee | Multimedia Information Group | Oct. 25, 2023

- Employed **abstract syntax tree** to **parse SQL** into various components and implemented a method for generating synthetic dataset with over 10,000 SQL based on adjustable semantic rules.
- Trained a large language model **BERT** on the synthetic dataset to automate **SQL segmentation and labeling**, treating it as a **Named Entity Recognition** task.
- Introduced syntax errors and semantic errors into the dataset as data augmentation, strengthening the fault tolerance capability of the system.

PROFESSIONAL EXPERIENCE

Northeastern University **Jan. 2022 ~ Dec. 2023**

Teaching Assistant

- Selected as a Teaching Assistant for **3 courses**: Object-Oriented Design, Computer Network and Algorithms.
- Conducted weekly office hours to address students' inquiries and provided assistance with their homework.
- Developed weekly assignments and laboratory exercises, evaluated homework, and graded exams.

The Commons XR **Jan. 2023 ~ Apr. 2023**

Data Engineer Intern

- **Led two individuals** from the product and data team to construct a metrics monitoring webpage (Azure-based). Developed to resolve a long-standing data inconsistency complained about by the data team.
- Designed the UI/UX using **TypeScript** & **React.js** and embedded Power BI to provide real-time dashboards on the front end. Used JWT in cookies to deliver personalized data and **Redis** for caching.
- Built robust **Restful APIs** with Nest.js. Leveraged **Spark** and Python to subsample data, reducing the data volume in SQL Server by **70%**, accelerating query speed, and greatly improving data team productivity.
- Utilized Stream Analytics to read data from the Event Hub. Employed windowing functions to subsample data from **30 to 0.5 msg/s**, dramatically reducing the workload of browsers and databases.

Sleep Number Lab **May 2022 ~ Aug. 2022**

Cloud Engineer Intern

- Assisted the Cloud team (AWS-based) in building a **distributed data platform** to capture **1 billion** daily bio-data using **Spring** and **Kafka**. Implemented a **real-time data pipeline** (~1s lag) to transform data from CSV to Parquet for the machine learning team.
- Optimized an Lambda function with multiprocessing techniques, resulting in a **30% cost reduction** and reducing the average processing time within the pipeline by **46%**.
- Utilized SQL-like queries to extract and store data into DynamoDB and S3, triggered through EventBridge.
- Migrated MQTT brokers from self-managed servers to SaaS and developed an **script for over 1 million IoT devices** using Python for seamless provision in IoT Core through the usage of Cognito and IAM.

RESEARCH INTERESTS

- Spatio-Temporal Data Mining, Deep Learning, Smart City, Intelligent Transportation System, IoT Network

SKILLS

Languages: Java, Python, Typescript, JavaScript, SQL, Go, C, HTML/CSS
Frames & Tools: Spring, Flask, gRPC, React, Node, Express, Nest, Kafka, MyBatis, Docker, AWS, Azure, GCP

SEMINARS, WORKSHOPS, TRAININGS

- Seminar: Privacy Preserving Deep Learning for IoT: Game Theoretical Model, Tapia Conference, Sep., 2023
- Seminar: Decarbonizing Scope 3 On-Road Transport Emissions, Verge Conference, Oct., 2023
- Seminar: Building a More Hopeful Climate Workforce, Verge Conference, Oct., 2023
- Seminar: Vision HGNN: An Image is More than a Graph of Nodes, MIG, Oct., 2023
- Workshop: Developing Large-Scale Parallel Programs in Python with Parsl, Tapia Conference, Sep., 2023
- Workshop: A Picture is Worth a Thousand Data Points: Intro to Visualization, Tapia Conference, Sep., 2023
- Traning: CS230: Deep Learning, Standford Online, summer 2022
- Traning: Urban Geography, Undergraduate Course, fall 2019
- Traning: The Principle of Geographic Information System, Undergraduate Course, fall 2019
- Training: Planning of Urban Transportation, Undergraduate Course, spring 2018