Qirui Wang

qw43@cs.washington.edu | (206) 849-2864 | github.com/Typhoeus-Wang

EDUCATION

University of Washington, Seattle **Seattle, WA**

- **B.S. in Computer Science + Applied Computational Mathematical Science** (Double Major) ***Sep 2020 - Jun 2024***
- **GPA:** 3.94/4 **Awards:** Annual Dean's Honor List
- **Courses:** Deep Learning, Linear Algebra, Data Structures, Differential Equation, Probability, Discrete Mathematics Modeling, Databases and Data Management, Statistical Methods for Data Science, Software Design

SKILLS SUMMARY

- **Languages:** Java, Python, Matlab, R, JavaScript, SQL, NoSQL
- **Frameworks and packages:** PyTorch, Numpy, Pandas, Matplotlib, Scipy, Scikit-Learn, tqdm, NodeJS, React
- **Models:** Transformer, BERT, ViT, word2vec, GloVe, CLIP, ResNet, Encoder-Decoder, LSTM, RNN, CNN
- **Statistics:** Bootstrapping, Monte Carlo Simulation, Significance Testing, Regression, AIC, BIC, EDF
- **Areas of Interests:** Speech Processing, Natural Language Processing, Multimodal Learning

RESEARCH EXPERIENCES

Spoken Language Systems, MIT — Remote

- **Distillation of Speech Self-Supevised Learning using Mamba** ***Feb 2024 - Present***
  - (In progress), with ***Alexander Liu*** and ***James Glass***

Mobile Intelligence Lab, University of Washington, Seattle — Seattle, WA

- **Real Time Spatial Speech Translation** ***Apr 2024 - Present***
  - (In Progress) with ***Shyam Gollakota*** and ***Luke Zettlemoyer***

- **Conversation Dataset and Benchmark** ***Aug 2024 - Present***
  - (In Progress) Working on **Conversation Dataset and Benchmark** project with ***Vidya Srinivas*** and ***Shyam Gollakota***

- **Target Conversation Extraction (InterSpeech 2024),** paper ***Aug 2024 - Present***
  - Target Conversation Extraction with ***Tuochao Chen*** and ***Shyam Gollakota***
  - Generated 20000 training data by cleaning, resampling and mixing several speech sources from conversation corpus
  - Built a whole training pipeline to automate model training, experiment, and metric value visualization
  - Designed a summarizer model using CNN, LSTM, and FiLM and to get speech embedding from clean speech example
  - Participated in model architecture design to incorporate speech embedding in multi-speaker speech separation

Lilian Ratliff's Group, University of Washington, Seattle — Seattle, WA

- **Effect of Adaptation Rate and Cost Display in a Human-AI Interaction Game** ***Jun 2022 - Jan 2024***
  - Discussed different human computer game settings including 2x2, 1x2, and 2x1with ***Lillian Ratliff***
  - Ran experiments and collected data with different learning rate from different game settings
  - Utilized regression algorithms to calculate **nash** equilibrium and **stackelberg** equilibrium and made visualization
  - Developed a website platform using FastAPI to facilitate participants to take experiments

Gemoetric Data Analysis, University of Washington, Seattle — Seattle, WA

- **Manifold Learning Examples,** github repo ***Jun 2022 - Dec 2022***
  - Explored different manifold learning algorithms and their limitations on high dimension data with ***Marina Meila***
  - Experimented manifold learning algorithms including **Isomap**, **Spectral Embedding**, **LLE**, **T-SNE**, **UMAP**
  - Ran these algorithms with parameters on datasets including rectangle, rectangle with a hole, torus and swiss-roll
  - Examined the embeddings of each dimension reduction algorithm and compiled the results into a repository

WORK EXPERIENCES

NetUp — Seattle, WA

- **Machine Learning Engineering Part-Time** ***Oct 2023 - Present***
  - Built a web scraper using beautifulsoup and scraped 10000 industry and company information for data generation
  - Lead and design a recommendation system (idea from TinVec) and tested it using synthetic user history data
  - Built an ETL to transfer user interaction data stored in **DynamoDB** and **EC2** to prepared periodic training dataset
  - Deployed recommendation system to **Sagemaker** and employed **Lambda** to trigger model update

USAFacts - Ballmer Group — Seattle, WA

- **Machine Learning Engineering Intern** ***Jun 2023 - Sep 2023***
  - Built a ChatGPT plugin to enable **Retrieval Argmented Generation** with government data through ChatGPT
  - Conducted research, drafted design document using **Confluence**, and wrote timeline and tickets on **Jira**

  - Developed API endpoints using **FastAPI** to upsert document and retrieve documents to vector database
  - Utilized **Cognitive Search** as vector database and stored vectorized documents using OpenAI's embedding model
  - Deployed API to **Azure** with **Docker** and used **Github actions** as **CI/CD** to support continuous deployment
  - Scraped text data over 1000 webpages using **BeautifulSoup** and deployed serverless function to **Azure Function**
  - Stored the scraped data in **PostgreSQL** and populate Cognitive Search vector database for QA in batches

TEACHING EXPERIENCES

University of Washington, Seattle — Seattle, WA

- **Teaching Assistant** at UW Paul G. Allen School of Computer Science and Engineering ***Sep 2022 - Present***
  - Teach SQL, Database Design, Cloud Database Application, NoSQL, Data Serialization for 5 quarters.
  - Holding quiz sections and office hours to teach students basic ideas of Database: data models, query languages, transactions, database tuning, and parallelism, and guided them with hands-on experience with Azure.
  - Grading students' assignments and exams and give their feedback.
  - Checking Ed message boards and email regularly and answering their questions about course content

PROJECT EXPERIENCES

- **Evaluation of Effect of Presentational Factors on Academic Paper Success** ***Aug 2023 - Dec 2023***
  - Cleaned Semantic Scholar dataset and preprocessed 4.3 million CS papers for readability and sentiment analysis
  - Scraped 200 influential CS papers using beautifulsoup to validate our finds about paper success
  - Defined academic success by quantifying each paper's citation count within five years of its publication, establishing a measurable standard for paper impact
  - Calculated readability using the FOG Index, determining the years of formal education needed for comprehension
  - Identified content-based presentational factors, including positive and argumentative language, through a combination
  - Visualized the relations between presentational factors and citation metric and found citation is parabolically correlated with argumentative languages.

- **Evaluation on Bird Classification with Unimodality versus Multimodality** ***Jun 2023***
  - Merged from multimodal bird data with overlapped species to obtain 64 classes with 11,435 images and 5,257 audio files
  - Preprocessed audio files to 2D Mel-Spectrograms and then transformed them to 3-channel tensor for training
  - Fine-tuned an image classifier, an audio classifier based on Resnet50 and achieved F1 score of 0.85 and 0.54 respectively
  - Mapped image and text to a same representation space using CLIP and fine-tuned OpenAI ViT-B-32 and a softmax layer
  - Evaluated two types of classifiers and found multimodal classifier perform better in bird classification

- **Multi-Label Text Classification using BERT PyTorch** ***Nov 2022***
  - Examined original toxic comments dataset and resampled 15,000 clean examples to obtain a balanced dataset
  - Wrapped tokenization process with BERT tokenizer in my customed dataset to facilitate training
  - Set up an optimizer scheduler to let it grow 0.001 per step during the warm-up and then go down (linearly) to 0
  - Combined BCELoss with a sigmoid loss to calculate the loss and set up area under ROC as evaluation metric
  - Fine-tuned a pre-trained BERT and a fully connected layer, achieved 98.13% accuracy and over 98% AUROC per class

- **Heart Attack Disease Analysis and Prediction** ***Oct 2022***
  - Preprocessed a heart attack disease dataset by removing null values to obtain 304 data samples
  - Conducted EDA using KDE to examine distribution of continuous data and contingency table for categorical data
  - Calculated correlation matrix to select features whose correlation score less than 0.7 for training
  - Selected best multinomial logistic regression which achieved 0.85 F1 score with 8 features using LOO cross-validation
  - Bootstrapped 10,000 samples to construct 95% confidence interval to estimate uncertainty of significant coefficients

- **Online Vaccine Scheduler (Azure, SQL)** ***Nov 2021***
  - Built an online vaccine appointment scheduler using SQL and Java
  - Utilized Azure to build database to store login information and vaccine, availability, appointment information
  - Programmed a command line user interface using java to navigate user how to search and reserve appointments