Youjia Li

3611 Washington St, Jamaica Plain
MA 02130, USA
(+1) 267 670 9164
lyoujia2021@gmail.com

Education

2017–2019 **University of Pennsylvania** *Philadelphia, PA*
Master of Science in Engineering: Data Science, GPA 4.0/4.0
- Advisor: Professor Zachary G. Ives
- Coursework in Mathematical Statistics, Data Mining, Applied Machine Learning, Big Data Analytics, Computer Vision, Deep Learning, Optimization, Database and Information System

2013–2017 **University of Nottingham** *Nottingham, UK*
Bachelor of Engineering: Electrical and Electronic Engineering, GPA 3.98/4.0
- Degree Classification: First Class

Research Experience

Jan 2019 - May 2019 **Distributed Computation of Mayo Clinic's Seizure Detection Challenge** *University of Pennsylvania*
Independent Study, Advisor: Professor Zachary G. Ives
- Implemented distributed seizure detection algorithm on Spark to predict stages of epilepsy (non-seizure, early-stage seizure, or seizure onset) based on EEG recordings collected from epilepsy patients to improve scalability and runtime performance of detection system
- Designed and implemented distributed data transformation and feature selection pipelines to pre-process EEG signals through FFT and correlation analysis to build feature vectors in time and frequency domain, which were piped into random forest classifiers built with Spark MLib
- Employed different sizes of clusters on GCP Dataproc to execute pipeline and reduced runtime of data processing and modeling by up to 51.7% with no-worse accuracy as compared to local run

Sept 2016 - May 2017 **Adaptive Equalization in Digital Communications** *University of Nottingham*
Bachelor Thesis, Advisor: Professor Malcolm Woolfson
- Developed two adaptive equalizers (decision-directed and decision-feedback) to compensate for time-varying channel effect and noise during digital communication
- Implemented adaptive equalizers in MATLAB by constructing adaptive digital filters with weights updated iteratively through error term between output and reference signal by gradient descent
- Evaluated equalizers' performances by signal waveform MSE reduction and decision errors reduction measured by BER (bit error ratio); decision-feedback scheme turns out to be superior with a 41.7% MSE reduction and 69.5% BER reduction
- Expanded scope of project by completing stretch goals - testing equalizer performance with modulated (phase-shift-keying) input signal

Teaching Experience

Sept 2018 - Dec 2018 **Graduate Teaching Assistant** *University of Pennsylvania*
CIS545 Big Data Analytics

Publications

[1]Li, X., Huang, S., Zhao, H., Guo, X., Xu, L., Li, X., & Li, Y. (2016). Image compression based on restricted wavelet synopses with maximum error bound. Proceedings of the 9th International Conference on Utility and Cloud Computing. http://doi.org/10.1145/2996890.3007880

Professional Experience

Aug 2019 - Present **Senior Data Scientist** *Wayfair* *Boston, MA*

- Designed and developed next generation pre-order incident forecasting model which predicts probability of major incident types and their corresponding financial impact by leveraging survival analysis and gradient boosting tree model, which would raise gross profit by 0.8% (~$3M in US markets per month)
- Collaborated with machine learning engineers to productionize developed machine learning models and migrate existing data products to GCP by leveraging PySpark, Docker and GCP AI platforms
- Built ETL pipeline and interactive dashboard to track daily predictions of Wayfair's major profitability model by leveraging Elasticsearch, Kibana, and Airflow
- Implemented and maintained expedited edition of Wayfair's profitability model for high volume periods including Black Friday and Cyber Monday
- Promoted to senior in 2020 winter cycle

May 2018 - Aug 2018 **Data Scientist Intern** *Knaq* *Jersey City, NJ*

- Engaged in developing prototype cloud-based predictive maintenance online platform for vertical transportation system (elevator) with focus on anomaly detection based on sensor signals
- Built online position detection system to determine elevator's real-time positions from previous trips using DBSCAN clustering and configured RabbitMQ channel to transmit predictions back to monitor dashboard

Project Experience

Feb 2019 - May 2019 **Gendered Pronoun Coreference Resolution with BERT** *University of Pennsylvania*

Deep Learning Course Project

- Built gender-invariant pronoun resolver to predict which entity name pronoun refers to out of two candidate names in Pytorch
- Leveraged BERT as pre-train encoder to generate embeddings and built initial model pipeline with fine-tuned BERT baseline model which achieved 3-class log-loss of 0.53 in cross validation
- Our final resolution with series connection architecture of BERT and Gated-RGCN ranked in top 8% of the Google AI GAP Kaggle competition (Bronze)

Oct 2018 - Dec 2018 **Moving Object Recognition and Tracking in Videos** *University of Pennsylvania*

Computer Vision Course Project

- Devised approaches to recognize and track objects in videos without having to exhaustively feed each frame to object detection NN model
- Parsed videos with OpenCV and employed pre-trained Mask R-CNN model to recognize objects from 81 possible classes and generate bounding boxes in first frames of videos
- Devised iterative image processing procedure to track objects in subsequent frames with KLT optical flow method based on inherited bounding boxes from first frame and identified moving objects in sliding cameras

Awards and Fellowships

2021 2nd Prize, Wayfair Pricing and Merch Hackathon

2019 Kaggle Competition Bronze Prize

2019 Analyst Fellows, Wharton Analytics Fellows Program

2018 Finalist, Citadel Data Open Princeton Datathon

2017 Nottingham Advantage Award, University of Nottingham

2014-2015 Head's Scholarship, Dean's Scholarship, University of Nottingham

Skills

Programming **Proficient:** Python, SQL(MySQL, SQLite), JavaScript **Basic:** C++, Java, , HTML, CSS

ML Random Forest, XGBoost, LightGBM, Neural Networks, SVM, K-means, DBSCAN

Libraries Scikit-learn, PySpark, Pandas, PyTorch, OpenCV, Plotly, Matplotlib

Software Jupyterlab, PyCharm, Git, Airflow, Google Cloud Platform, Docker, DBeaver, DataGrip, Tableau, LaTeX

Certificates

Cousera Deep Learning Specialization by DeepLearningAI, Big Data Specialization by UCSD